UNIVERSIDADE FEDERAL DO PARANÁ - UFPR
PROGRAMA DE PÓS-GRADUAÇÃO EM MÉTODOS NUMÉRICOS EM ENGENHARIA - PPGMNE

ROSANGELA VILLWOCK

TÉCNICAS DE AGRUPAMENTO E DE HIERARQUIZAÇÃO NO CONTEXTO DE *KDD* – APLICAÇÃO A DADOS TEMPORAIS DE INSTRUMENTAÇÃO GEOTÉCNICA-ESTRUTURAL DA USINA HIDRELÉTRICA DE ITAIPU

CURITIBA
2009

ROSANGELA VILLWOCK

TÉCNICAS DE AGRUPAMENTO E DE HIERARQUIZAÇÃO NO CONTEXTO DE *KDD* – APLICAÇÃO A DADOS TEMPORAIS DE INSTRUMENTAÇÃO GEOTÉCNICA-ESTRUTURAL DA USINA HIDRELÉTRICA DE ITAIPU

Tese apresentada ao Programa de Pós-Graduação em Métodos Numéricos em Engenharia, Área de Concentração de Programação Matemática, dos Setores de Tecnologia e de Ciências Exatas, da Universidade Federal do Paraná, como requisito parcial à obtenção do título de Doutor em Ciências.

Orientadora: Profª. Dra. Maria Teresinha Arns Steiner
Co-orientadores: Profª. Dra. Andréa Sell Dyminski
Prof. Dr. Paulo Henrique Siqueira

CURITIBA

2009

TERMO DE APROVAÇÃO

ROSANGELA VILLWOCK

TÉCNICAS DE AGRUPAMENTO E DE HIERARQUIZAÇÃO NO CONTEXTO DE *KDD* – APLICAÇÃO A DADOS TEMPORAIS DE INSTRUMENTAÇÃO GEOTÉCNICA-ESTRUTURAL DA USINA HIDRELÉTRICA DE ITAIPU

Tese aprovada como requisito parcial para obtenção do grau de Doutor em Ciências, no Programa de Pós-Graduação em Métodos Numéricos em Engenharia – Programação Matemática da Universidade Federal do Paraná, pela seguinte banca examinadora:

Orientadora: _____
Prof.ª Dr.ª Maria Teresinha Arns Steiner
Programa de Pós-Graduação em Métodos Numéricos em Engenharia e Coordenação de Engenharia de Produção, UFPR

_____
Prof.ª Dr.ª Andréa Sell Dyminski
Programa de Pós-Graduação em Métodos Numéricos em Engenharia e Departamento de Construção Civil, UFPR

_____
Prof. Dr. Paulo Henrique Siqueira
Programa de Pós-Graduação em Métodos Numéricos em Engenharia e Departamento de Expressão Gráfica, UFPR

_____
Prof. Dr. Anselmo Chaves Neto
Programa de Pós-Graduação em Métodos Numéricos em Engenharia e Departamento de Estatística, UFPR

_____
Prof. Dr. Celso Romanel
Departamento de Engenharia Civil, PUC–RJ

_____
Prof. Dr. Júlio Cesar Nievola
Programa de Pós-Graduação em Informática Aplicada, PUC–PR

Curitiba, 29 de julho de 2009.

Aos meus pais Luiz e Neusa.

Às minhas orientadoras Maria Teresinha e Maria Hermínia.

AGRADECIMENTOS

À Deus, pela vida, bênção e proteção.

À Professora Dra. Maria Teresinha Arns Steiner, pela orientação para a realização deste trabalho, pelo apoio e pelo incentivo em todas as fases do curso.

À Professora Dra. Andréa Sell Dyminski, pela co-orientação deste trabalho, pelo apoio e pelo incentivo.

Ao Professor Dr. Paulo Henrique Siqueira, pela co-orientação deste trabalho, pelo apoio e pelo incentivo.

À minha família, pelo apoio, pelo incentivo e pela paciência durante a realização deste curso.

Às minhas tias Raquel e Suely, pelo apoio e pelo carinho.

À Bernadete Maria Suaki Brandão, pela amizade e pelo apoio.

Aos Professores do Programa de Pós-Graduação em Métodos Numéricos em Engenharia, pelos ensinamentos transmitidos.

Ao Professor Dr. Julio Cesar Nievola, pelos ensinamentos transmitidos.

Ao Professor Dr. Leandro dos Santos Coelho, pelas valiosas sugestões.

Ao Professor Dr. Anselmo Chaves Neto, pelas valiosas sugestões.

Ao Professor Dr. Wagner M. N. Zola, pela disponibilização do recurso da grade computacional do LCPAD: Laboratório Central de Processamento de Alto Desempenho/UFPR, para execução deste trabalho.

À equipe de engenheiros da Itaipu, pelos dados de instrumentação e contribuições técnicas.

À equipe do projeto AIVEC - Análise de Incertezas e Estimação de Valores de Controle para o Sistema de Monitoração Geotécnico-estrutural na Barragem de Itaipu, pela colaboração.

Aos colegas de Pós-graduação, pela amizade, convívio e apoio.

Ao José Henrique Ferreira, pela disposição em ajudar quando necessário.

À Maristela Bandil, pela motivação e eficiência nos serviços prestados.

À Universidade Federal do Paraná, pela oportunidade de realização deste curso.

À Universidade Estadual do Oeste do Paraná, pela licença concedida.

À FINEP, pelo apoio financeiro ao projeto de pesquisa AIEVC - Análise de Incertezas e Estimação de Valores de Controle para o Sistema de Monitoração Geotécnico-estrutural na Barragem de Itaipu e ao projeto CT – INFRA / UFPR / Modelagem e Computação Científica.

À CAPES, pelo auxílio financeiro.

A todos, que de alguma forma contribuíram para a realização deste trabalho.

RESUMO

O monitoramento da estrutura de uma barragem, de importância bem conhecida, pode gerar uma enorme massa de dados, definidos em domínios multidimensionais, cuja análise e interpretação nem sempre são triviais. É importante selecionar as informações que melhor "expliquem" o comportamento da barragem, permitindo a previsão e a resolução de eventuais problemas que possam ocorrer. A Usina Hidrelétrica de Itaipu, maior geradora de hidroeletricidade do mundo, possui mais de 2.200 instrumentos que monitoram seu comportamento geotécnico e estrutural, os quais possuem leituras armazenadas em um banco de dados há mais de 30 anos. Assim sendo, o objetivo principal deste trabalho é apresentar uma metodologia, enquadrada na área de *KDD,* "Descoberta de Conhecimento em Bases de Dados", com o intuito de realizar a hierarquização de instrumentos de monitoramento de barragens, maximizando a eficácia e eficiência das análises das leituras, através da identificação de grupos de instrumentos semelhantes e, também, detectando os principais instrumentos. A metodologia foi aplicada à 30 extensômetros localizados em diferentes blocos do trecho F da referida barragem que, com uma, duas ou três hastes, totalizam 72 medidas de deslocamentos, das quais 24 foram automatizadas pela empresa. Para a fase de pré-processamento dos dados, do processo *KDD*, identificou-se que para a maioria dos instrumentos tem-se uma leitura mensal, porém, alguns deles, apresentam mais de uma leitura por mês sendo que, nestes casos, foi obtida a média mensal. Por outro lado, alguns instrumentos apresentaram leituras faltantes e, nestas situações, foram realizadas interpolações por séries temporais garantindo, desta forma, que todos os instrumentos tivessem exatamente 120 leituras (10 anos). Já para a fase de Mineração de Dados, do processo *KDD*, a tarefa é o agrupamento de padrões e, para isso, foram utilizados os seguintes métodos: da área de Análise Estatística Multivariada (Ligação Simples, Média, Completa e Método Ward); da área de Redes Neurais Artificiais (Redes Neurais de Kohonen) e da área de Metaheurísticas (foi proposto um Algoritmo de Agrupamento Baseado em Formigas). Em relação ao algoritmo proposto, este foi testado em três bases de dados reais (IRIS, WINE e PIMA Indians Diabetes) e em duas bases de dados reais de séries temporais (GUN e LIGHTNING-2), sendo que o seu desempenho foi comparado com o de outros dois métodos (Método Ward e Redes Neurais de Kohonen). Na aplicação da Análise de Agrupamento (pelo Método Ward) aos dados de instrumentação geotécnica-estrutural da Itaipu, mostrou-se que é possível encontrar justificativas técnicas para a formação dos grupos, inclusive identificando um grupo de hastes de maior importância. Já a aplicação da Análise Fatorial aos referidos dados, mostrou-se bastante eficaz para realizar a hierarquização das hastes de extensômetros, com base nas comunalidades. No algoritmo proposto, as principais modificações em relação ao algoritmo básico proposto por Deneubourg *et al.* (1991, *apud* Handl, Knowles e Dorigo, 2006), foram: a introdução de uma comparação da probabilidade de descarregar um padrão na posição escolhida aleatoriamente com a probabilidade de descarregar este padrão em sua posição atual; a introdução de uma avaliação da probabilidade de uma posição vizinha, quando a decisão de descarregar um padrão for positiva e a célula em que o padrão deveria ser descarregado estiver ocupada; e a substituição do padrão carregado por uma formiga, caso este padrão não seja descarregado em 100 iterações consecutivas. O algoritmo proposto apresentou resultados satisfatórios em relação aos resultados de Boryczka (2008) para as bases de dados reais e, quando aplicado aos dados de instrumentação geotécnica-estrutural da Itaipu, o mesmo foi capaz de identificar o grupo de hastes de maior importância.

ABSTRACT

The monitoring of the dam structures, of known importance, can generate an enormous mass of data, defined in multidimensional domains, which analysis and interpretation are not trivial. It is important to select the information that best "explains" the behavior of the dam, allowing the forecast and the resolution of eventual problems that can happen. The Hydroelectric Power Plant of Itaipu, the largest hydro electrical power producer of the world, has more than 2.200 instruments to monitor its geotechnical and structural behavior, which has readings stored in a database for more than 30 years. In this way, the main goal of this work is to present a methodology, framed in the KDD area*, "Knowledge Discovery in Databases"*, in order to carry out the ranking of instruments of monitoring of dams, maximizing the effectiveness and the efficiency of the readings analyses, through the identification of groups of similar instruments and, also, detecting the main instruments. The methodology was applied to 30 extensometers located in different blocks of the sector F of the referred dam which, with one, two or three rod, totalized 72 measures of displacements, of which 24 were automated by the company. For the phase of preprocessing of the data, of the KDD process, it was identified that the majority of the instruments had a monthly reading, however, some of them, presented more than a reading by month and, in these cases, it was obtained the monthly average. In the other hand, some instruments presented failed readings and, in these situations, interpolations were carried out by time series assuring, in this way, that all of the instruments had exactly 120 readings (10 years). In the Data Mining phase, of the process KDD, the task is to group the patterns and, for that, the following methods were used: of the Multivariate Statistical Analysis area (Single Linkage, Average Linkage, Complete Linkage and Ward Method); of the Artificial Neural Networks area (Kohonen Maps) and of the Metaheuristics area (it was proposed an Ant Based Clustering Algorithm). In relation to the proposed algorithm, it was tested in three real databases (IRIS, WINE and PIMA Indians Diabetes) and in two time series real databases (GUN and LIGHTNING-2), and their performances were compared with other two methods (Ward Method and Kohonen Maps). In the application of the Clustering Analysis (by Method Ward) at the data of instrumentation geotechnical and structural of the Itaipu, it was shown that it is possible to find technical justification for the formation of the groups and, also, identifying a group of rods of greatest importance. The application of the Factorial Analysis to the referred data showed to be effective to realize the extensometer rods ranking, based in the communality. In the proposed algorithm, the main modifications in relation to the basic algorithm proposed by Deneubourg *et al.* (1991, *apud* Handl, Knowles e Dorigo, 2006), were: the introduction of a comparison of the probability of drop a pattern in a random chosen position with the probability of drop this pattern in the current position; the introduction of a evaluation of the probability of a neighboring position, when the decision of dropping a pattern is positive and the cell in which the pattern should be dropped is busied; and the replacement of the carried pattern by an ant, in case this pattern is not dropped in 100 consecutive iterations. The proposed algorithm presented satisfactory results compared with Boryczka (2008)'s results for the real databases and, when applied to the data of instrumentation geotechnical and structural of the Itaipu, the same was able to identify the group of rods of greatest importance.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA SIGLAS

A^2CA – Adaptative Ant Clustering Algoritthm – Algoritmo de Agrupamento por Formigas Adaptável

ACAM – Ant-based clustering algorithm – Algoritmo de Agrupamento Baseado em Formigas Modificado

ACO – Ant Colony Optimization – Otimização por Colônia de Formigas

ACS – Ant Colony System – Sistema de Colônia de Formigas

AS – Ant System – Sistema de Formiga

CBGB – Comitê Brasileiro de Grandes Barragens

DM – Data Mining – Mineração de Dados

ICOLD - International Comission on Large Dams – Comissão Internacional de Grandes Barragens

KDD – Knowledge Discovery in Databases – Descoberta de Conhecimento em Bases de Dados

SOM – Self Organizing Map – Mapas auto-organizáveis

TSP – Traveling Salesman Problem – Problema do Caixeiro Viajante

LISTA DE SÍMBOLOS

α – porcentagem de padrões na grade classificados como semelhantes

η – taxa de aprendizagem

ρ – matriz de correlação

Σ – matriz de covariância

σ – raio de vizinhança ou percepção

ε_i – i-ésimo erro ou fator específico

$\underline{\mu}$ – média ou valor esperado

ψ_i – variância específica

AIC – critério de Akaike

D – Índice Dunn

d(*i*, *j*) – dissimilaridade ou distância entre *i* e *j*

$E(\underline{X})$ – valor esperado de $\underline{X}$

F_j – j-ésimo fator comum

f(i) – função de vizinhança

h_i^2 – comunalidade

ℓ_{ij} – peso do *j*-ésimo fator F_j na *i*-ésima variável X_i

N – número máximo de iterações

N_{occ} – número de células da grade ocupadas

P – peso próprio

P_{drop} – Probabilidade de descarregar padrões

P_{pick} – Probabilidade de carregar padrões

Q – matriz dos dados padronizada

R – Índice Aleatório

Sim – medida de similaridade

SQE – soma do quadrado do erro

$V(\underline{X})$ – variância de $\underline{X}$

$\underline{w_j}$ – peso sináptico do neurônio j

$\underline{X}$ – vetor aleatório

SUMÁRIO

1 INTRODUÇÃO

1.1 O PROBLEMA

Uma vez que os potenciais prejuízos e riscos decorrentes de acidentes em barragens podem assumir grandes dimensões, um projeto seguro, uma construção adequada e a correta operação de barragens são preocupações de âmbito mundial. Além disso, um efetivo monitoramento em grandes barragens é imprescindível para a segurança de sua estrutura. Diretrizes internacionais visando a segurança de barragens e muitas discussões produtivas sobre este tema têm sido propostas e conduzidas, tais como a da Comissão Internacional de Grandes Barragens (*ICOLD - International Comission on Large Dams*) (ICOLD, 2008). No Brasil, diretrizes visando à segurança de barragens foram publicadas pelo Comitê Brasileiro de Grandes Barragens em 1983 (CBGB, 1983). Além disso, a Comissão de Constituição e Justiça e de Cidadania aprovou no dia 23/06/2009 a proposta que obriga o Poder Executivo a instituir uma Política Nacional de Segurança de Barragens. O seu objetivo foi dotar o Poder Público de um instrumento permanente de fiscalização das mais de 300 mil barragens existentes no País. O texto acatado é o substitutivo ao Projeto de Lei 1181/03. A proposta original, Projeto de Lei – PL 1181/03 – BRASIL (2003), de autoria do deputado Leonardo Monteiro, define diretrizes de segurança para construção de barragens de água e de aterros para contenção de resíduos líquidos industriais.

Exemplos recentes de rupturas de barragens no Brasil podem ser citados: a ruptura da barragem de Câmara, PA, em 2004; a ruptura da estrutura de desvio da barragem Campos Novos, SC, em 2006; a ruptura da barragem Algodões I, PI, em 2009; dentre outros.

Segundo Kalustyan (1999), as catástrofes têm sido sinais oportunos para a inspeção de critérios de projeto existentes e seleção de métodos mais efetivos de monitoramento da segurança de barragens.

Yenigun e Erkek (2007) apresentam uma tabela contendo estimativas das causas mais comuns de rupturas em barragens, dentre as quais destacam-se as seguintes: problemas de fundação; vertedouro inadequado; problemas de construção; recalques diferenciais; subpressão elevada; ruptura de aterros; materiais

defeituosos; operação incorreta; atos de guerra e terremotos. Todos estes problemas podem ser diagnosticados com o monitoramento da instrumentação da barragem, com exceção dos dois últimos, cujas frequências percentuais somam apenas 4%.

Segundo Menescal (2009), a experiência mundial mostra que os custos para garantir a segurança de uma barragem são pequenos quando comparados aos custos em caso de ruptura. O autor ainda comenta sobre a importância da utilização de um banco de dados de instrumentação para subsidiar a análise preliminar das leituras, detectando anomalias.

O monitoramento da estrutura de uma barragem, de importância bem conhecida, pode gerar uma enorme massa de dados, definidos em domínios multidimensionais, cuja análise e interpretação nem sempre são triviais. É importante selecionar as informações que melhor “entendam” o comportamento da barragem, permitindo a previsão e a resolução de eventuais problemas que possam ocorrer.

Uma interessante discussão sobre a avaliação de risco e de tomada de decisão para a segurança de barragens é apresentada em Bowles *et al.* (2003). Este artigo propõe uma matriz de justificativa e recomendação de decisão. As avaliações propostas são adaptáveis a qualquer prática atual de engenharia de barragens, avaliação de risco de segurança de barragens e outros fatores de decisão. A abordagem pode ser útil em três tipos de decisão: estabelecer metas de risco toleráveis; identificar um caminho de redução de riscos e administrar o risco residual.

Harrald *et al.* (2004) fazem uma revisão sobre alguns sistemas e metodologias para a tomada de decisões com o intuito de auxiliar na priorização de tarefas e diminuição do risco de falhas. Entre os sistemas e as metodologias apresentadas no artigo estão a Metodologia de Avaliação de Risco para Barragens, a Metodologia de Avaliação de Risco Portfólio, o Sistema Modelo Baseado em Risco para Segurança de Barragens, o Índice de Condição, entre outros. Os autores apresentam ainda uma matriz de comparação dos métodos de análise de risco para segurança de barragens.

A Usina Hidrelétrica de Itaipu, maior geradora de hidroeletricidade do mundo, possui mais de 2.200 instrumentos que monitoram seu comportamento geotécnico e estrutural, os quais possuem leituras armazenadas em um banco de dados há mais de 30 anos. A alta dimensionalidade e a grande quantidade de

registros contidos nas bases de dados são problemas não triviais tendo-se em vista a busca pelo “conhecimento” a partir destes dados.

Este trabalho apresenta três principais contribuições, dentre outras consideradas secundárias, sendo ainda abordado um importante problema de engenharia, a análise de dados de instrumentação de grandes obras.

A primeira contribuição diz respeito à aplicação de técnicas de agrupamento, dentre outras, no contexto de *KDD*, do inglês *“Knowledge Discovery in Databases”* ou “Descoberta de Conhecimento em Bases de Dados”, tendo como objetivo a identificação dos instrumentos que são realmente significativos à análise do comportamento de uma barragem.

As novas propostas apresentadas ao algoritmo de Agrupamento baseado em Colônia de Formigas formam a segunda grande contribuição deste trabalho. Esta metaheurística, relativamente nova, ainda exige muita investigação para melhorar seu desempenho.

A terceira contribuição foi a aplicação deste algoritmo proposto a bases de dados de séries temporais. Poucos algoritmos de agrupamentos, recentemente criados, têm sido utilizados no agrupamento de séries temporais. Neste trabalho, os métodos de agrupamento foram aplicados diretamente às bases de dados de séries temporais, sem a aplicação de um método de pré-processamento dos dados visando o agrupamento dos dados especificamente para séries temporais.

1.2 OBJETIVOS

1.2.1 Objetivo Geral

O objetivo principal deste trabalho é apresentar uma metodologia, enquadrada na área de *KDD*, com o intuito de realizar o agrupamento e a hierarquização de instrumentos de monitoramento de barragens, maximizando a eficácia e a eficiência das análises das leituras, através da identificação de grupos de instrumentos semelhantes e, também, detectando os principais instrumentos.

Vale salientar que tal metodologia, que envolve a proposta de um Algoritmo para Agrupamento baseado em Colônias de Formigas, poderá ser utilizada não apenas aos dados de monitoramento de outras barragens, mas, também, a quaisquer outras bases de dados.

1.2.2 Objetivos Específicos

a) Aplicar a Análise de Agrupamento, através da Análise Multivariada, em cinco bases de dados reais e de séries temporais.
b) Aplicar a Análise de Agrupamento, através das Redes Neurais de Kohonen Unidimensional, às cinco bases de dados reais e de séries temporais.
c) Propor alterações e melhorias no algoritmo de Agrupamento baseado em Formigas originalmente proposto por Deneubourg *et al.* (1991, *apud* Handl, Knowles e Dorigo, 2006).
d) Aplicar a Análise de Agrupamento, através do algoritmo de Agrupamento baseado em Formigas proposto, às cinco bases de dados reais e de séries temporais, comparando os resultados obtidos com outros dois métodos de agrupamento.
e) Aplicar a Análise de Agrupamento, através do algoritmo de Agrupamento baseado em Formigas proposto, aos dados de instrumentação geotécnica-estrutural da Usina Hidrelétrica de Itaipu, mais especificamente, aos instrumentos chamados extensômetros, localizadas no trecho F da barragem.
f) Aplicar a Análise Fatorial aos extensômetros de cada grupo de instrumentos, fazendo a hierarquização dos mesmos.

1.3 JUSTIFICATIVA

A análise minuciosa dos dados dos instrumentos de auscultação exige a combinação de conhecimentos de engenharia com matemática e estatística, bem como experiência prévia do engenheiro ou técnico responsável pela interpretação destes dados, consumindo muito tempo e, muitas vezes, inviabilizando o cumprimento desta tarefa de forma eficiente. Por esta razão, o uso de técnicas e ferramentas computacionais, que auxiliem o tomador de decisões, é extremamente relevante.

Muitas vezes, um grande volume de dados contém informações úteis, as quais pode-se chamar de “conhecimento”, sendo que, em geral, esta informação não está facilmente disponível ou identificada. Analistas humanos podem gastar

semanas para descobrir este conhecimento e, por este motivo, alguns bancos de dados grandes nunca recebem uma análise detalhada adequada como deveriam (TAN, STEINBACH; KUMAR, 2005). Na medida em que há o aumento da quantidade de dados há, também, o aumento da aplicação de técnicas de Mineração de Dados. Dados inteligentemente analisados constituem um valioso recurso para a tomada de decisões (WITTEN; FRANK, 2000).

Além disso, não há registros da existência de métodos que realizem a hierarquização de instrumentos de monitoramento em barragens. Em caso de necessidade de intensificação de leituras, esta hierarquização poderia ser utilizado para definir, dentre os instrumentos, aqueles que seriam os escolhidos.

Ainda, ao repetir-se o procedimento de análise dos instrumentos em períodos subseqüentes, a mudança na situação de um instrumento indicaria a necessidade de uma investigação mais aprofundada no mesmo.

1.4 ESTRUTURA DO TRABALHO

O texto está organizado da seguinte forma:

No capítulo 2 é apresentada uma revisão bibliográfica sobre Segurança de Barragens e Instrumentos de Monitoramento, explicitando a necessidade e a importância da instrumentação para garantir a segurança. Também é apresentada a descrição do Processo *KDD*, das tarefas de Mineração de Dados (agrupamento) e dos métodos que serão utilizados neste trabalho, sempre relacionando-os com trabalhos já apresentados na literatura. Os métodos utilizados para o agrupamento e descritos neste capítulo 2, são: técnicas da área de Análise Multivariada dos Dados; Redes Neurais de Kohonen e Agrupamento baseado em Formigas e, finalmente, são apresentadas técnicas para a avaliação de agrupamentos.

No capítulo 3 são apresentadas seis bases de dados, com as quais se trabalhou: reais (3); de séries temporais (2) e a dos dados de instrumentação da barragem de Itaipu. As primeiras cinco bases de dados foram utilizadas com o intuito de melhor compreender as técnicas apresentadas na literatura, permitindo, então, o desenvolvimento de novas contribuições, como as já citadas.

Ainda neste capítulo 3 é apresentada a 1ª. fase do processo *KDD* (etapas de seleção, pré-processamento e formatação dos dados) aplicada aos dados de instrumentação de Itaipu. Em seguida, é apresentada a maneira como os métodos

de agrupamento para a Mineração dos Dados (2ª. fase do processo *KDD*), Análise Estatística Multivariada e Redes Neurais de Kohonen serão aplicadas às bases de dados. E, finalmente, as principais contribuições (modificações e melhorias) para o Agrupamento baseado em Colônias de Formigas são descritas.

No capítulo 4 são apresentados os resultados, discussões e figuras ilustrativas sobre a aplicação da proposta para o Agrupamento baseado em Formigas, nas bases de dados reais e de séries temporais, bem como o seu desempenho quando comparado aos outros dois métodos aqui abordados (da Análise Multivariada e Redes Neurais de Kohonen). Também são apresentados os resultados da aplicação do algoritmo proposto aos dados de instrumentação geotécnica-estrutural da Itaipu.

Finalmente, no capítulo 5, são apresentadas as conclusões e as sugestões para trabalhos futuros.

2. REVISÃO DE LITERATURA

Neste capítulo são abordados os diversos temas tratados aqui neste trabalho (Segurança de Barragens e o Processo *KDD*, do qual são destacadas as seguintes etapas: Análise Estatística Multivariada; Redes Neurais de Kohonen; Metaheurística para Agrupamento baseada em Formigas; Agrupamento em Séries Temporais e Avaliação de Agrupamento), assim como diversas referências relacionadas aos mesmos.

2.1 A SEGURANÇA DE BARRAGENS

O conceito de "Segurança de Barragens" envolve aspectos estruturais, hidráulicos, geotécnicos, ambientais e operacionais. Estas características devem ser consideradas durante toda a vida útil da barragem. Um sistema de instrumentação capaz de monitorar o comportamento geotécnico e estrutural de uma barragem é essencial para avaliar seu comportamento e integridade. Uma boa revisão sobre a importância da instrumentação para a avaliação da segurança de uma barragem pode ser encontrada em Dibiagio (2000) e Duarte, Calcina e Galván (2006).

Alguns objetivos da instrumentação de barragens e sua relação com segurança estrutural são descritos em dois Manuais de Engenharia publicados por U.S. Army Corp de Engenheiros (1987 e 1995). Nestes manuais, os principais objetivos de um plano de instrumentação geotécnico são agrupados em quatro categorias: avaliação analítica; predição de desempenho futuro; avaliação jurídica, desenvolvimento e verificação de projetos futuros. A instrumentação pode alcançar estes objetivos provendo dados quantitativos para acessar informações úteis como pressão piezométrica, deformação, tensão total e níveis de água. Com inspeções visuais e periódicas combinadas com análise de dados cuidadosa uma condição crítica pode ser revelada (FEMA, 2004).

Para Saré *et al.* (2006), o monitoramento de barragens assume diferentes características e finalidades dependendo da etapa da obra que se deseja analisar. Ao longo de sua vida útil, podem-se detectar variações nas condições de segurança. Para Duarte, Calcina e Galván (2006), a instalação de um sistema de instrumentação geotécnica é uma das medidas mais importantes e necessárias para garantir um nível de segurança adequado para uma barragem. Este sistema permite

acompanhar o nível de segurança durante a vida útil da obra, verificando se tudo se mantém dentro das premissas estabelecidas em projeto.

A necessidade de construção de novas barragens, a preocupação com a recuperação e manutenção de barragens já existentes, o fato de muitas destas obras não receberam os cuidados necessários à sua manutenção, motivaram Menescal (2009) a apresentar uma proposta de organização institucional e de procedimentos, a fim de dotar o Brasil com um Sistema Integrado de Gestão da Segurança de Barragens. Segundo este autor somente com um grande esforço de melhoria da gestão da segurança, as barragens poderão atender às necessidades da população, sem representarem fonte de riscos permanentes.

Segundo Krüger (2008), a construção de barragens é um exemplo da obrigatoriedade da consideração dos aspectos de incerteza e risco para a obtenção de uma estrutura com desempenho e segurança adequados. No Brasil, apesar do grande número de barragens construídas e projetos em andamento, os critérios de projeto são permanentemente questionados e revisados. No trabalho deste autor, o principal objetivo foi desenvolver uma metodologia para a análise de confiabilidade estrutural de barragens de concreto. Foram desenvolvidas equações de estado limite para vários modos de falha em barragens de concreto à gravidade: flutuação, tombamento, deslizamento e tensões normais. O procedimento desenvolvido foi testado e validado a partir de dados do concreto compactado com rolo (CCR) da barragem de Salto Caxias, situada no Rio Iguaçu, Estado do Paraná.

Os princípios estabelecidos na NBR 8681 – Ações e Segurança das Estruturas (ABNT, 2003) conceituam a segurança das obras de concreto de uma barragem. Em projetos de barragens de concreto à gravidade, são necessárias verificações correspondentes à análise de estabilidade, visando avaliar a segurança quanto aos movimentos: deslizamento, tombamento, flutuação, tensões na base da fundação e na estrutura, deformações, recalques e vibrações.

A estabilidade da barragem deve ser primeiramente analisada durante a fase de projeto. A geometria das estruturas e as propriedades dos materiais envolvidos devem ser consideradas bem como as condições de carregamento. Algumas condições de carregamento básicas são mostradas na figura 2.1.

Fisicamente se explica que a diferença de nível de água (montante-jusante) gera um gradiente hidráulico entre montante e jusante da barragem, fazendo com que a água do reservatório queira passar para jusante buscando o equilíbrio

hidráulico. Para tal, a água percola através do maciço de fundação da barragem. Durante este processo, a água infiltrada gera forças verticais que atuam de baixo para cima sob a barragem, denominadas subpressões na fundação. Representa-se por $F_{subpressão}$ a resultante destas forças. Além disso, água do reservatório gera forças horizontais que atuam de montante à jusante sobre a barragem, denominadas pressões hidrostáticas contra a parede da barragem. Representa-se por $F_{reservatório}$ a resultante destas forças. Estas duas forças resultantes são chamadas forças desestabilizadoras. Já a força P (o peso próprio da barragem) é uma força estabilizadora da estrutura. A combinação $F_{subpressão}$ e de $F_{reservatório}$ pode gerar o tombamento e/ou deslizamento da barragem, tanto pelos esforços e momentos diretamente aplicados quanto pelo alívio do peso próprio da estrutura (no caso das subpressões).

Os efeitos das cargas na barragem, acima descritos, podem ser observados na figura 2.1, onde são enfatizados o deslizamento (a) e o tombamento (b).

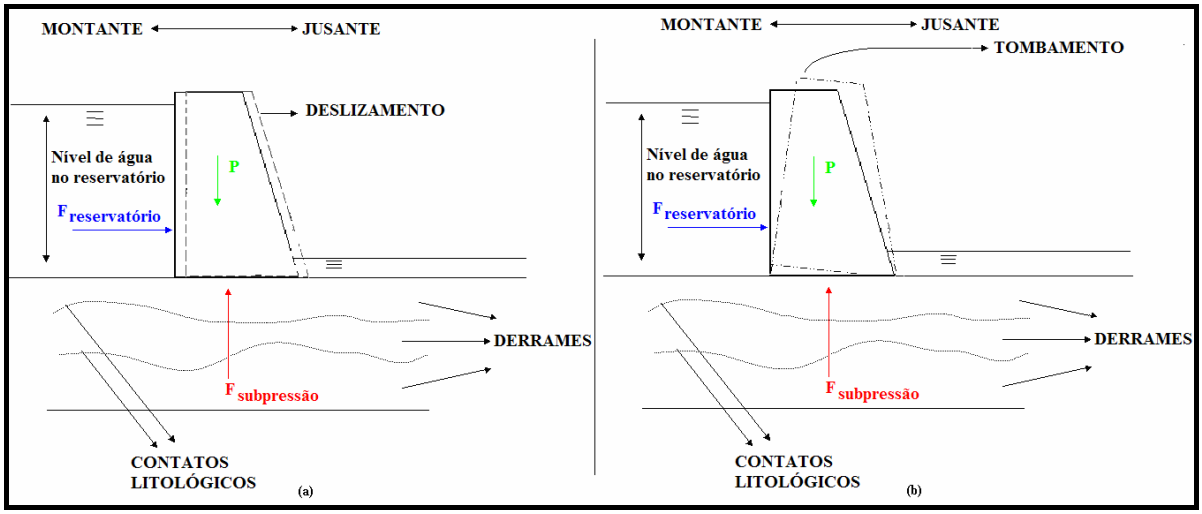

Figura 2.1 – Condições de carga básicas e formas de instabilização de barragens de gravidade de concreto.

As condições de carregamento e propriedades de materiais podem mudar ao longo do ciclo de vida da barragem e a instrumentação pode identificar algumas destas mudanças.

A figura 2.2 mostra as diferenças no comportamento da barragem quanto às condições climáticas características de verão e inverno, bem como suas conseqüências. No verão ocorre uma dilatação no concreto, o que provoca o tombamento do bloco à montante. Este tombamento, por sua vez, faz com que o

bloco comprima a fundação. No inverno o concreto se contrai, provocando um tombamento do bloco à jusante, voltando a posição inicial. Isto faz com que a pressão exercida sobre a fundação, ocorrida no verão, seja aliviada. Pode-se então identificar um comportamento cíclico da estrutura, intimamente condicionado a condições ambientais que envolvem a obra.

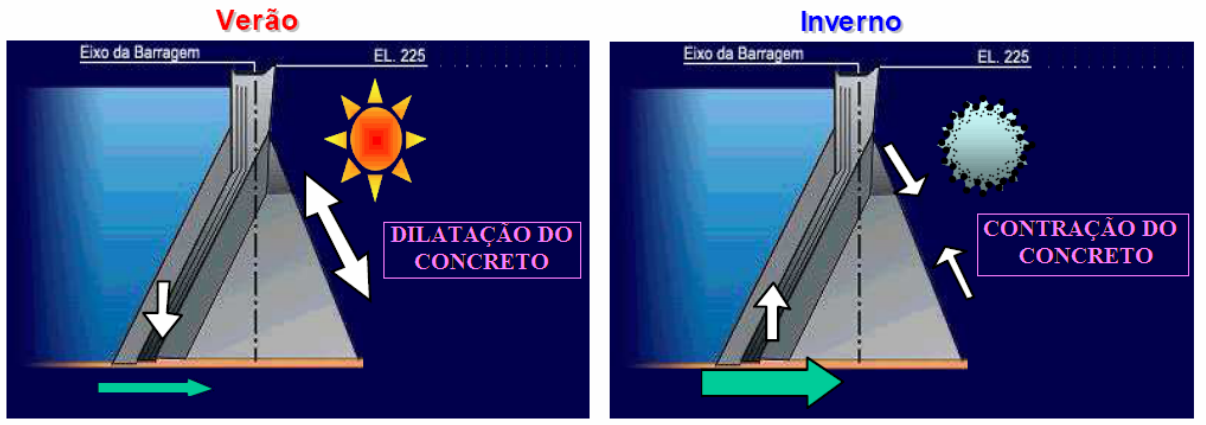

Figura 2.2 – Comportamento da barragem com relação às condições climáticas características de verão e inverno (Adaptada de Osako, 2002).

2.1.1 Os Instrumentos de Monitoramento

Segundo a FEMA (2004), a instrumentação deve ser usada como suplemento às inspeções visuais na avaliação do desempenho e da segurança das barragens. A inspeção cuidadosa dos dados de instrumentação pode revelar uma condição crítica.

A figura 2.3 apresenta as correlações entre os tipos de instrumentos usualmente empregados na auscultação de barragens de concreto e os principais tipos de deterioração de barragens de concreto (SILVEIRA, 2003). Observando-se esta figura, o extensômetro múltiplo, por exemplo, está relacionado com o monitoramento de deterioração por escorregamento, recalque diferencial, subsidência do terreno, distenção no pé de montante e reatividade Álcali-Agregado.

A medição dos recalques de uma barragem de concreto é uma das observações mais importantes na supervisão do comportamento da estrutura durante os períodos de construção, enchimento do reservatório e operação da barragem. A medição de recalque pode ser realizada por extensômetros múltiplos de hastes instalados em furos de sondagem (SILVEIRA, 2003). A figura 2.4 mostra o

extensômetro múltiplo de haste e um exemplo de um perfil típico de um extensômetro múltiplo de haste na Itaipu.

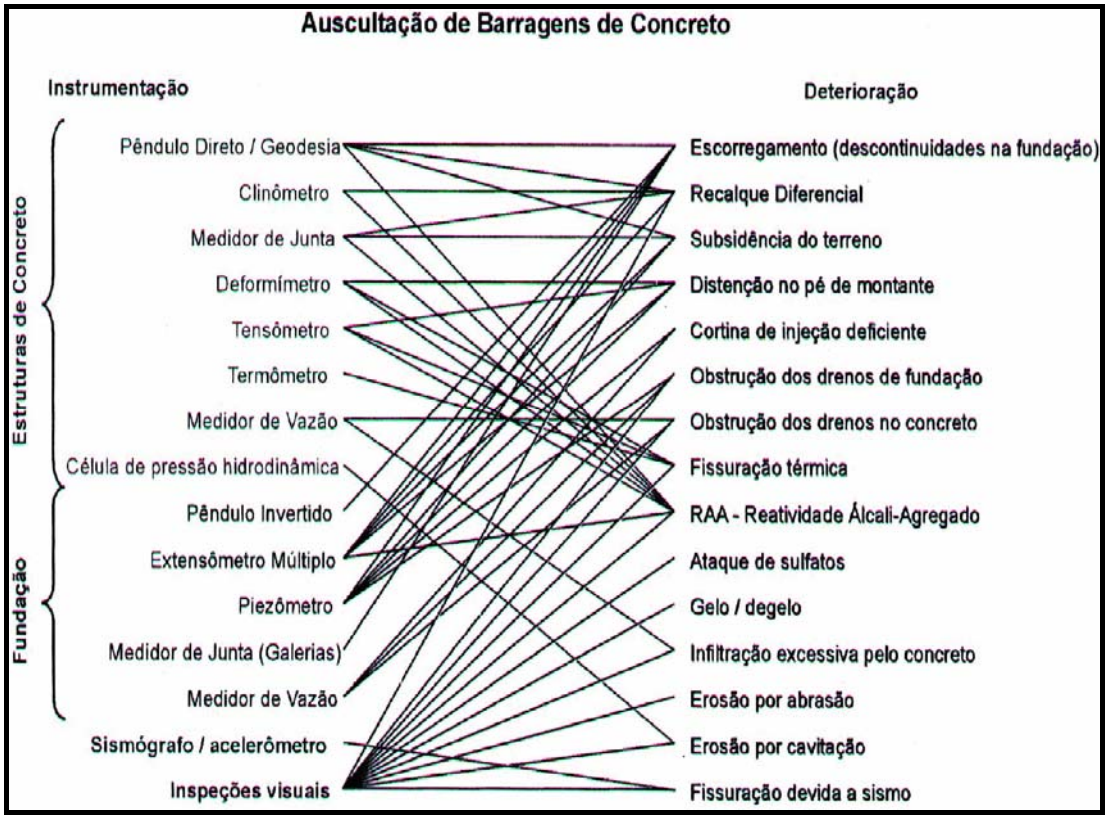

Figura 2.3 – Correlação entre os tipos de instrumentos e a deterioração de barragens de concreto (SILVEIRA, 2003).

Com o uso de várias hastes, pode-se fazer a medição dos deslocamentos e deformações em vários trechos da fundação, dentre eles, contato concreto-rocha, juntas-falhas e outras descontinuidades subhorizontais na fundação. Esta abordagem foi usada na barragem de Itaipu, onde diversos pontos do maciço de fundação foram instrumentados, em especial as descontinuidades geológicas. A figura 2.5 mostra um perfil geológico típico do maciço de fundação do trecho sem túnel da Barragem Lateral Direita da Itaipu, onde podem-se observar as principais descontinuidades (contatos, brechas e juntas) daquele sitio. Nos blocos onde há galerias de acesso transversais ao eixo (como na figura 2.6), a instalação de extensômetros a montante e a jusante permite medir deslocamentos angulares da barragem junto à fundação (SILVEIRA, 2003).

A medição de deslocamentos horizontais da crista são parâmetros de relevante importância, afetados por deflexões da estrutura de concreto, rotação da

base da estrutura (devido à deformabilidade da fundação) ou influências térmicas ambientais. Estes deslocamentos são afetados por características do concreto ou por propriedades do maciço rochoso de fundação, resultando em importantes informações para a auscultação do comportamento da barragem e de sua fundação. Os deslocamentos horizontais da crista podem ser medidos por pêndulos diretos, normalmente instalados ao final do processo construtivo. As medições ocorrem nas fases de enchimento do reservatório e operação da barragem (SILVEIRA, 2003).

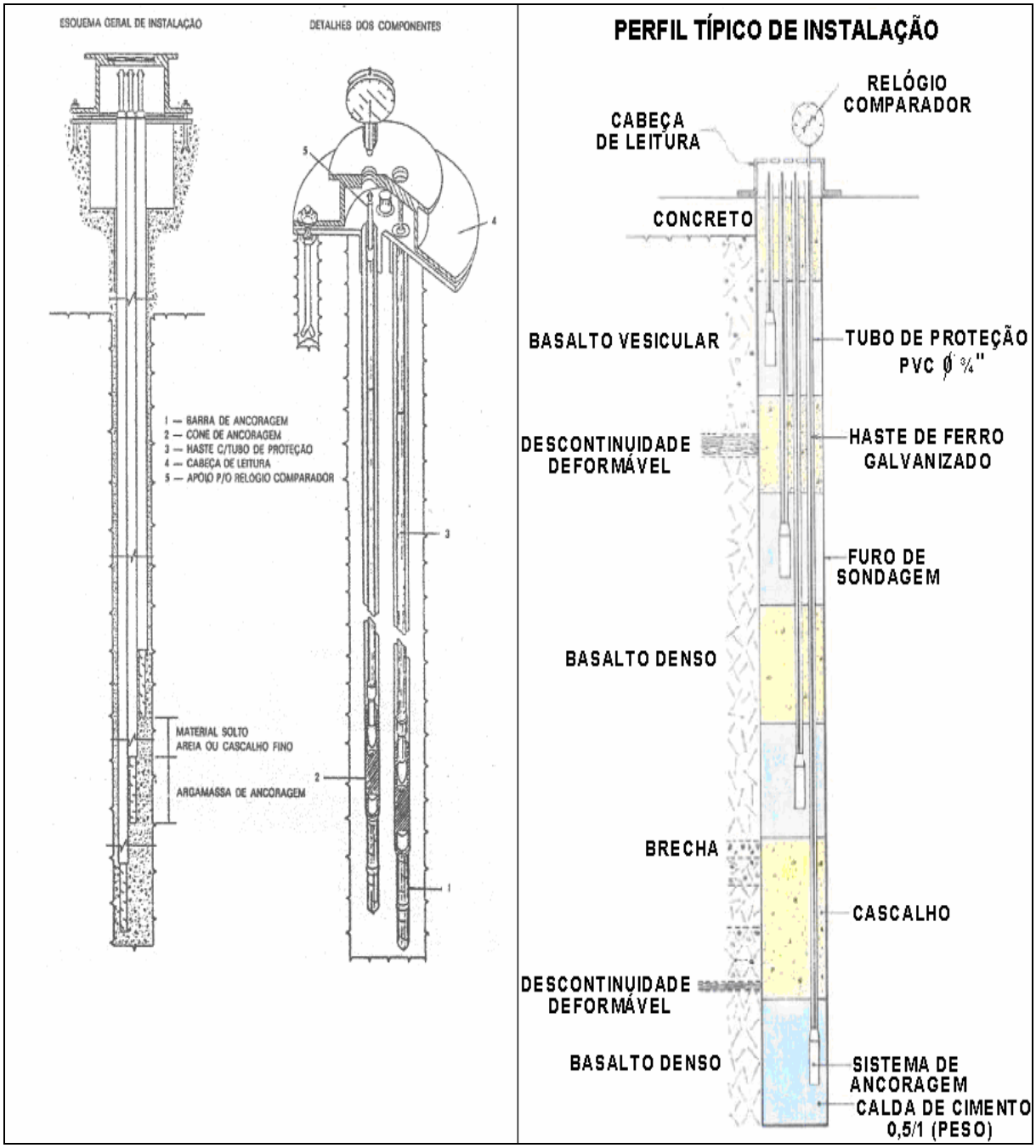

Figura 2.4 – Extensômetro múltiplo de haste e um exemplo de um perfil típico de um extensômetro múltiplo de haste na Itaipu (Adaptada de MATOS, 2002).

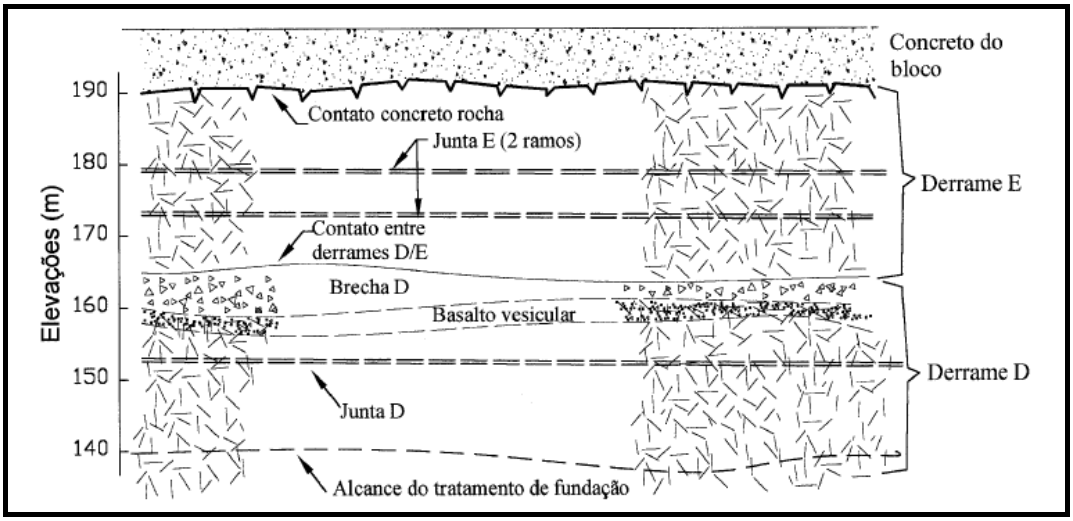

Figura 2.5 – Perfil geológico esquemático da fundação da Itaipu (ITAIPU BINACIONAL, 1995, *apud* OSAKO, 2002).

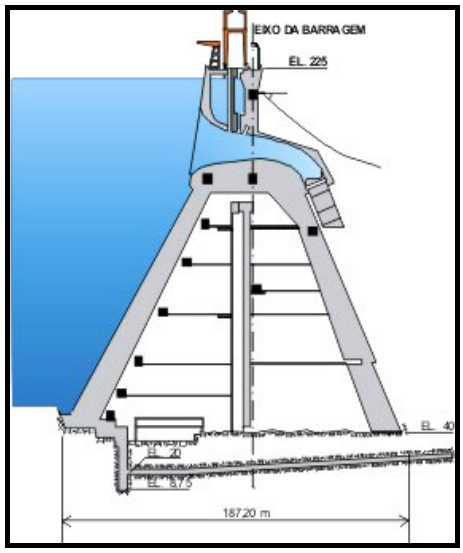

Figura 2.6 – Blocos com galeria de acesso transversal ao eixo (ITAIPU, 2008).

A estabilidade da estrutura em termos de escorregamento, tombamento ou flutuação é diretamente afetada pelo nível das pressões piezométricas na interface concreto-rocha e nas descontinuidades subhorizontais de baixa resistência existentes na fundação. A medição das subpressões na fundação das barragens de concreto é importante para a supervisão de suas condições de segurança. A drenagem é um dos meios mais eficientes para assegurar coeficientes de segurança adequados. As medidas de subpressões são realizadas pelos piezômetros (SILVEIRA, 2003).

Outra medição importante é a das vazões de drenagem pela fundação e infiltrações, através do concreto de uma barragem. Estas medições devem ser feitas

durante o enchimento do reservatório e no período de operação, pois refletem imediatamente muitos dos problemas que ocorrem com esse tipo de estrutura. As medições podem ser realizadas com a instalação de medidores de vazão do tipo triangular ao longo das canaletas de drenagem, junto ao piso das galerias (SILVEIRA, 2003).

2.2 O PROCESSO *KDD*

Segundo Fayyad *et al.* (1996), o processo *KDD,* do inglês *"Knowledge Discovery in Databases"* ou "Descoberta de Conhecimento em Bases de Dados", é um processo não trivial de descoberta de padrões válidos, novos, úteis e acessíveis. A principal vantagem do processo de descoberta é que não são necessárias hipóteses, sendo que o conhecimento é extraído dos dados sem conhecimento prévio.

Muitas vezes a expressão "Mineração de Dados" (do inglês "*Data Mining*" – *DM*) é usada como sinônimo do processo *KDD*. Segundo Diniz e Louzada-Neto (2000), a mineração de dados é uma parte do processo *KDD* que se relaciona com a análise de dados e o uso de ferramentas computacionais na busca de padrões (característica, regras e regularidades) em um grande conjunto de dados.

O processo *KDD* é um conjunto de atividades contínuas que são compostas, basicamente, por cinco etapas: seleção dos dados, pré-processamento, formatação ou transformação, Mineração de Dados e interpretação dos resultados, como ilustrado na figura 2.7.

Primeiramente deve-se ter domínio da aplicação e objetivos claros. Na primeira etapa são selecionados e coletados os dados necessários. Na etapa de pré-processamento verificam-se os dados faltantes ou inconsistentes. Na etapa de transformação há uma preparação dos dados visando à aplicação da Mineração de Dados, usando métodos de redução de dimensionalidade dos dados, por exemplo. A etapa de Mineração de Dados é o núcleo do processo, onde são aplicados os algoritmos para extrair padrões dos dados. A etapa de Interpretação dos resultados consiste em validar o conhecimento extraído (FAYYAD *et al.*, 1996). Segundo Silver (1996), as etapas de pré-processamento e formatação podem levar até 80% do tempo necessário de todo o processo.

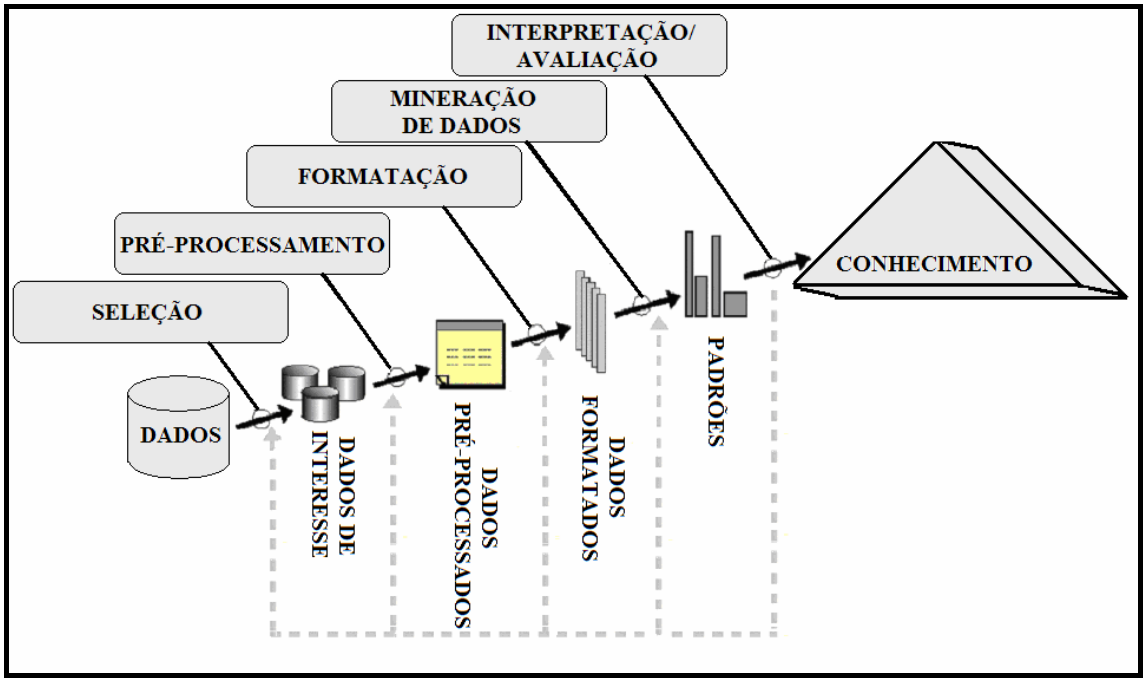

Figura 2.7 – Etapas do processo *KDD*, adaptada de Fayyad *et al.* (1996).

Dentro do contexto do processo *KDD*, alguns termos são muito usados e estão bem definidos em Witten e Frank (2000). São eles:

- Conceito: resultado do processo de aprendizado;
- Exemplos (também chamados de instâncias ou padrões): entradas do processo de aprendizagem, um conjunto de características que definem um objeto;
- Atributos (também chamados de características): qualquer medição útil extraída no processo de identificação do padrão; podem ser nominais ou numéricos, contínuos ou discretos.

O principal objetivo do processo *KDD* é extrair o conhecimento a partir de informações “escondidas” nos dados que sejam úteis nas tomadas de decisões, utilizando métodos, algoritmos e técnicas de diferentes áreas científicas, que segundo Tan, Steinbach e Kumar (2005) incluem Estatística, Inteligência Artificial, Aprendizagem de Máquinas e Reconhecimento de Padrões.

2.2.1 Tarefas e Métodos de Mineração de Dados

As tarefas de Mineração de Dados podem ser preditivas ou descritivas. As preditivas usam algumas variáveis para prever valores desconhecidos ou futuros de

outras variáveis, enquanto que as descritivas encontram padrões para descrever os dados. As principais tarefas de Mineração de Dados estão relacionadas à Classificação, Associação e Agrupamento de padrões (FAYYAD *et al.*, 1996).

Na Classificação, cada padrão contém um conjunto de atributos e um dos atributos é denominado classe. O objetivo da classificação é encontrar um modelo para predição da classe como função dos outros atributos (TAN; STEINBACH; KUMAR, 2005). A regressão é um caso particular da classificação, já que seu objetivo é encontrar um modelo para predição de um atributo contínuo como função dos outros atributos.

Já na Associação, o objetivo é produzir regras de dependência que irão predizer a ocorrência de um atributo baseado na ocorrência de outros atributos (TAN; STEINBACH; KUMAR, 2005). Regras de associação não são diferentes de regras de classificação, exceto pelo fato de que elas podem predizer algum atributo, não necessariamente a classe (WITTEN; FRANK, 2000).

O Agrupamento ou Segmentação (*Clustering*) procura grupos de padrões tal que padrões pertencentes a um mesmo grupo são mais similares uns aos outros e dissimilares a padrões em outros grupos. Segundo Hair Jr *et al.* (2005), a análise de agrupamentos é uma técnica analítica para desenvolver subgrupos significativos de objetos. Seu objetivo é classificar os objetos em um pequeno número de grupos mutuamente excludentes. Para Freitas (2002), na análise de agrupamento é importante favorecer um número pequeno de grupos.

2.2.1.1 A Tarefa de Agrupamento

Os algoritmos de agrupamento podem ser divididos em categorias de diversas formas de acordo com as características. As duas principais classes de algoritmos de agrupamento são: os métodos hierárquicos e os métodos de particionamento.

Os métodos hierárquicos englobam técnicas que buscam de forma hierárquica os grupos e, por isso, admitem obter vários níveis de agrupamento. Os métodos hierárquicos podem ser subdivididos em divisivos ou aglomerativos. O método hierárquico aglomerativo considera, a princípio, cada padrão como um grupo e, iterativamente, agrupa o par de grupos com maior similaridade em um novo grupo até formar um único grupo contendo todos os padrões. O método hierárquico

divisivo, ao contrário, inicia com um único grupo e executa um processo de sucessivas subdivisões (DINIZ; LOUZADA-NETO, 2000).

Os métodos de agrupamento hierárquicos mais populares são: Ligação Simples, Ligação Completa, Ligação Média e Método Ward. A forma mais comum de representar um agrupamento hierárquico utiliza um dendrograma, que representa o agrupamento dos padrões e os níveis de similaridade em que os grupos se formam. O dendrograma pode ser “quebrado” em diferentes níveis, mostrando diferentes grupos (JAIN; MURTY; FLYNN, 1999). No dendrograma da figura 2.8, admitindo um corte no nível apresentado na referida figura, observam-se dois grupos, sendo o primeiro composto pelos padrões *P1*, *P2* e *P5* e o segundo composto pelos padrões *P3* e *P4*.

Métodos não-hierárquicos ou de particionamento procuram uma partição sem a necessidade de associações hierárquicas. Seleciona-se uma partição dos elementos em *k* grupos, otimizando algum critério (DINIZ; LOUZADA-NETO, 2000).

O método mais conhecido entre os métodos de particionamento é o das *k*-médias (JOHNSON; WICHERN, 1998). Normalmente os *k* grupos encontrados são de melhor qualidade do que os *k* grupos produzidos pelos métodos hierárquicos. Os métodos de particionamento são vantajosos em aplicações que envolvem grandes séries de dados.

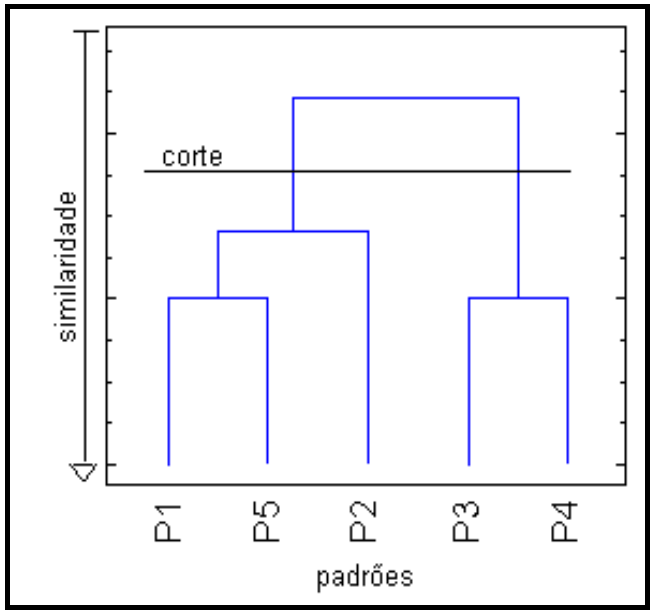

Figura 2.8 – Exemplo de dendrograma.

Outra característica importante que classifica algoritmos de agrupamento é a utilização de grades. Os métodos de agrupamento baseados em grade têm como principal característica a subdivisão do espaço em células. São exemplos de agrupamentos baseados em grade o Algoritmo de Agrupamento baseado em Colônia de Formigas e as Redes Neurais de Kohonen (KOHONEN, 1995).

Neste presente trabalho, a tarefa de Mineração de Dados está relacionada ao agrupamento de padrões. Os métodos a serem utilizados para realizar esta tarefa são: da área de Estatística Multivariada (Ligação Simples, Média, Completa e Método Ward), da área de Redes Neurais Artificiais (Redes Neurais de Kohonen Unidimensional) e da área de Metaheurísticas (Agrupamento baseado em Colônia de Formigas).

Os métodos da área de Estatística Multivariada foram utilizados por serem métodos já consagrados. A Análise Estatística Multivariada é um método antigo mas que foi viabilizado mais recentemente com a computação barata. As Redes Neurais de Kohonen Unidimensional foram utilizadas porque, assim como o Agrupamento baseado em Formigas, executam as tarefas de agrupamento e mapeamento topográfico, simultaneamente.

Além disso, o algoritmo de Agrupamento baseado em Colônia de Formigas foi escolhido para estudo, análise e novas propostas, devido a diversos fatores. Primeiramente, é uma metaheurística relativamente nova e tem recebido atenção especial, principalmente porque ainda exige muita investigação para melhorar seu desempenho, estabilidade e outras características, consideradas “chaves”, que fariam de tal algoritmo uma ferramenta madura para mineração de dados (BORYCZKA, 2009). Ainda, o referido algoritmo “consegue descobrir”, automaticamente, a quantidade de grupos nos padrões. Esta é uma vantagem, principalmente, na aplicação aos dados de instrumentação geotécnica-estrutural da Itaipu, pois não há conhecimento prévio que indique a quantidade de grupos.

Vale salientar que estes algoritmos que fazem mapeamento topográfico vão além de um mero agrupamento. Segundo Handl, Knowles e Dorigo (2006), eles não são limitados à descoberta de grupos homogêneos nos dados, mas também capturam relações de vizinhança numa visualização bi-dimensional de um espaço de dados de alta dimensão.

Outra questão, observada por Liao (2005), é que são poucos os estudos de agrupamentos relacionados a séries temporais que utilizam algoritmos de

agrupamento criados mais recentemente como, por exemplo, o Algoritmo de Agrupamento baseado em Colônia de Formigas. Vale ressaltar que neste estudo, a tarefa de agrupamento foi aplicada aos dados de instrumentação geotécnica-estrutural da Itaipu, que são séries temporais, além de outras bases de dados conforme será visto mais adiante.

Segundo Handl e Meyer (2007), o agrupamento com algoritmos baseados em enxames (*Swarm*) está emergindo como uma alternativa aos métodos mais convencionais, tais como o agrupamento hierárquico e o *k*-médias. Destes, o agrupamento baseado em formigas destaca-se como o mais utilizado grupo de algoritmos de agrupamento baseado em enxames.

2.3 ANÁLISE ESTATÍSTICA MULTIVARIADA

2.3.1 Análise Fatorial

A Análise Fatorial é um método estatístico cujo objetivo é explicar as correlações entre um conjunto grande de variáveis em termos de um conjunto de poucas variáveis aleatórias não-observáveis chamadas fatores. Assim, seja o vetor aleatório $\underline{X}$ composto por *p* variáveis aleatórias, $\underline{X}' = \begin{bmatrix} x_1 & x_2 & x_3 & \dots & x_p \end{bmatrix}$ e se deseja estudar a estrutura de covariância desse vetor, ou seja, se $\underline{X}$ for observado *n* vezes tem-se que os seus parâmetros $E(\underline{X}) = \underline{\mu}$ e $V(\underline{X}) = \Sigma$ podem ser estimados e o relacionamento entre as variáveis representado pela matriz de covariância Σ ou de correlação ρ avaliadas. A análise fatorial faz um agrupamento de variáveis para explicar a influência de variáveis latentes (não observáveis) ou fatores. Dentro de um mesmo grupo, as variáveis são altamente correlacionadas entre si, sendo que de um grupo para outro, as correlações são baixas. Cada grupo representa um fator, o qual é responsável pelas correlações observadas.

A matriz de covariância do vetor $\underline{X}$ pode ser colocada na forma exata: $V(\underline{X}) = \Sigma = LL' + \psi$, onde a matriz *LL'* tem na diagonal principal as chamadas comunalidades definidas para cada variável considerando-se *m* fatores por: $h_i^2 = \ell_{i1}^2 + \ell_{i2}^2 + \dots + \ell_{m}^2$ com i = 1, 2, ..., *p*. Assim, a comunalidade h_i^2 é a parte da variância da variável aleatória x_i que vem dos *m* fatores. E, a parte da variância da

variável aleatória x_i devida aos fatores p - m não importantes chama-se variância específica. Então, $V(x_i) = h_i^2 + \psi_i$.

Existem vários critérios para definir o número m de fatores. O critério mais utilizado é o critério de Kaiser (JOHNSON; WICHERN, 1998), que diz que o número de fatores extraídos deve ser igual ao número de autovalores maiores do que um.

Seja $\underline{X}$ um vetor aleatório, com p componentes, média $\underline{\mu}$ e matriz de covariância Σ. No modelo fatorial, $\underline{X}$ é linearmente dependente sobre algumas variáveis aleatórias não-observáveis F_1, F_2, ... , F_m chamadas fatores comuns e p fontes de variações aditivas: ε_1, ε_2, ... , ε_p, chamadas erros ou fatores específicos.

O modelo de Análise Fatorial é obtido pelo sistema abaixo, onde μ_i é a média da i-ésima variável, ε_i é o i-ésimo erro ou fator específico, F_j é o j-ésimo fator comum e ℓ_{ij} é o peso do j-ésimo fator F_j na i-ésima variável X_i. A equação 2.1 mostra o modelo na forma matricial.

$$\begin{cases} X_1 - \mu_1 = \ell_{11}F_1 + \ell_{12}F_2 + \ldots + \ell_{1m}F_m + \varepsilon_1 & i = 1,2,\ldots,p \\ X_2 - \mu_2 = \ell_{21}F_1 + \ell_{22}F_2 + \ldots + \ell_{2m}F_m + \varepsilon_2 & j = 1,2,\ldots,m \\ \ldots \\ X_p - \mu_p = \ell_{p1}F_1 + \ell_{p2}F_2 + \ldots + \ell_{pm}F_m + \varepsilon_p & m \leq p. \end{cases}$$

$$\underline{X} = \underline{\mu} + LL' + \psi \qquad (2.1)$$

Nesse modelo ortogonal assume-se que $E(\underline{F}) = \underline{0}_{m \times 1}$, $V(\underline{F}) = E(\underline{F}\,\underline{F}') = I_{m \times m}$, $E(\underline{\varepsilon}) = \underline{0}_{p \times 1}$, $V(\underline{\varepsilon}) = E(\underline{\varepsilon}\underline{\varepsilon}') = \Psi_{p \times p}$ (matriz diagonal com Ψs na diagonal) e $Cov(\underline{\varepsilon}, \underline{F}) = 0_{p \times m}$.

Segundo Johnson e Wichern (1998), o modelo e as restrições anteriores constituem o Modelo Fatorial Ortogonal.

Para estimar os pesos ℓ_{ij} e as variâncias específicas ψ_i, pode-se utilizar o método das componentes principais, que é descrito resumidamente a seguir (JOHNSON; WICHERN, 1998).

Sejam os pares de autovalores-autovetores (λ_i, $\underline{e}_i$) da matriz de covariância amostral S, com $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$. Seja $m < p$ o número de fatores comuns. A

matriz dos pesos estimados dos fatores é dada por $L = CD^{1/2}$, onde C é a matriz dos autovetores e D é uma matriz diagonal cujos elementos diagonais são os autovalores.

Na aplicação desse método, as observações primeiramente são centralizadas ou padronizadas. Neste caso, a matriz de covariância amostral S é a matriz correlação amostral R. As variâncias específicas ψ_i estimadas são fornecidas pelos elementos diagonais da matriz $\psi = S - LL'$.

Em muitas aplicações é preciso estimar o valor de cada um dos fatores (não observáveis) para uma observação individual $\underline{X}$, sendo que esses valores dos fatores são chamados de escores fatoriais. Os escores fatoriais estimados para as variáveis originais são $\underline{F} = (L'L)^{-1}L'(\underline{X} - \overline{X})$ e para as variáveis padronizadas são $\underline{F} = (L'L)'Lz$, desde que se use componentes principais para estimar os pesos.

Segundo Johnson e Wichern (1998), com a rotação dos fatores se obtém uma estrutura para os pesos tal que cada variável tenha peso alto em um único fator e pesos baixos ou moderados nos demais fatores. Kaiser sugeriu uma medida analítica conhecida como critério *Varimax* (JOHNSON; WICHERN, 1998).

Define-se por $\tilde{\ell}_{ij} = \frac{\ell_{ij}}{h_{ij}}$, os coeficientes rotacionados escalonados pela raiz quadrada das comunalidades. O procedimento *Varimax* seleciona a transformação ortogonal T que torna V (dado pela equação 2.2) o maior possível, ou seja, o procedimento parte de $\Sigma = LTT'L'$ e fornece os pesos $\underline{\ell}^*$ vindos de LT. Então, o critério é maximizar V.

$$V = \frac{1}{p}\sum_{j=1}^{m}\left[\sum_{i=1}^{p}\underline{\ell}^{*4}_{ij} - \left(\sum_{i=1}^{p}\underline{\ell}^{*2}_{ij}\right)/p\right] \qquad (2.2)$$

2.3.2 Análise de Agrupamento

O agrupamento de padrões é feito com base numa medida de similaridade ou dissimilaridade. A medida de similaridade avalia se os objetos são similares, ou seja, quanto maior o valor da medida mais parecidos são os objetos. A mais conhecida medida de similaridade é o coeficiente de correlação. A medida de

dissimilaridade avalia se os objetos são dissimilares, ou seja, quanto maior o valor da medida menos parecidos serão os objetos. A mais conhecida medida de dissimilaridade é a distância Euclidiana.

Na seção 2.2.1.1 foi citado que os métodos hierárquicos fazem agrupamentos ou divisões (nos métodos de agrupamento hierárquicos divisivos) iterativas de pares de grupos. Estes agrupamentos ou estas divisões são feitos por meio de ligações. Os tipos de ligações mais comuns são: Ligação Simples, Ligação Completa, Ligação Média e Método Ward, conforme já comentado também.

Na ligação simples (ou vizinho mais próximo), a distância entre dois grupos é a mínima das distâncias entre todos os pares de padrões *i* e *j*, *i* pertencente ao primeiro grupo e *j* ao segundo (JAIN; MURTY; FLYNN, 1999). Por exemplo, se um grupo 1 é formado pelos padrões *U* e *V* e um grupo 2 é formado pelo padrão *W*, a distância entre os grupos 1 e 2 é calculada $d_{(1,2)} = min\{d_{UW}, d_{VW}\}$ (JOHNSON; WICHERN, 1998).

Na ligação completa (ou vizinho mais distante), a distância entre dois grupos é a máxima das distâncias entre todos os pares de padrões *i* e *j*, *i* pertencente ao primeiro grupo e *j* ao segundo (JAIN; MURTY; FLYNN, 1999). Por exemplo, se um grupo 1 é formado pelos padrões *U* e *V* e um grupo 2 é formado pelo padrão *W*, a distância entre os grupos 1 e 2 é calculada $d_{(1,2)} = máx\{d_{UW}, d_{VW}\}$ (JOHNSON; WICHERN, 1998).

Já na ligação média, a distância entre dois grupos é a média das distâncias entre todos os pares de padrões, sendo que cada padrão do par é de um grupo. Se um grupo 1 é formado pelos elementos *U* e *V* e um grupo 2 é formado pelo elemento *W*, a distância entre os grupos 1 e 2 é calculada $d_{(1,2)} = \Sigma\Sigma\, d_{ik} / N_1 * N_2$, onde d_{ik} é a distância entre o padrão *i* no grupo 1 e o padrão *k* no grupo 2, N_1 é o número de padrões no grupo 1 e N_2 é o número de padrões no grupo 2 (JOHNSON; WICHERN, 1998).

Ainda segundo Johnson e Wichern (1998), o Método Ward faz a junção de dois grupos baseando-se na “perda de informação”. Considera-se como critério de “perda de informação” a soma do quadrado do erro (*SQE*). Para cada grupo *i*, calcula-se a média (ou centróide) do grupo e a soma do quadrado do erro do grupo *i* (SQE_i) que é a soma do quadrado do erro de cada padrão do grupo em relação à média. Para *k* grupos têm-se $SQE_1, SQE_2, ..., SQE_k$, onde define-se *SQE* pela equação 2.3.

$$SQE = SQE_1 + SQE_2 + ... + SQE_k \quad (2.3)$$

Para cada par de grupos *m* e *n*, primeiramente, calcula-se a média (ou centróide) do grupo formado (grupo *mn*). Em seguida, calcula-se a soma do quadrado do erro do grupo *mn* (SQE_{mn}), segundo a equação 2.4.

$$SQE = SQE_1 + SQE_2 + ... + SQE_k - SQE_m - SQE_n + SQE_{mn} \quad (2.4)$$

Os grupos *m* e *n* que apresentarem o menor aumento na soma do erro quadrático (*SQE*) (menor "perda de informação") serão unidos. Segundo Hair Jr *et al.* (2005), este método tende a obter grupos de mesmo tamanho devido a minimização de sua variação interna.

2. 4 REDES NEURAIS DE KOHONEN

Segundo Fausett (1994), Teuvo Kohonen, em 1982, desenvolveu o método de mapas de característica auto-organizáveis que faz uso de uma estrutura topológica para agrupar as unidades (padrões). *"Self Organizing Map"* (*SOM*; ou "Mapas auto-organizáveis"), também conhecidos por Redes Neurais de Kohonen, formam uma classe de Redes Neurais Artificiais em que a aprendizagem é não supervisionada.

Segundo Haykin (2001) o principal objetivo das Redes Neurais de Kohonen é transformar padrões de entrada de dimensão arbitrária em um mapa discreto. Os neurônios são colocados nos nós de uma grade, que pode ter qualquer dimensionalidade, normalmente são utilizadas grades bidimensionais (chamado de 2D-SOM). Existem ainda o 1D-SOM e o 3D-SOM, que utilizam grades (ou mapas) de uma e três dimensões, respectivamente.

Segundo Fausett (1994), as Redes Neurais Auto-Organizáveis são conhecidas por preservar a topologia. Segundo a autora, esta propriedade é observada no cérebro, mas não é encontrado em outras redes neurais artificiais.

O processo de aprendizagem de uma Rede Neural de Kohonen é baseado no aprendizado competitivo, onde os neurônios de saída da grade competem entre si para serem ativados. O neurônio de saída que vence a competição é chamado de neurônio vencedor. Todos os neurônios da grade devem ser expostos a um número

suficiente de padrões de entrada para assegurar o amadurecimento apropriado do processo de auto-organização (HAYKIN, 2001).

Segundo Haykin (2001), além do processo de competição, ainda são essenciais os processos de cooperação e adaptação para a formação do mapa. No processo de cooperação, o neurônio vencedor localiza o centro de uma vizinhança topológica de neurônios cooperativos. Para que o processo de auto-organização ocorra, no processo de adaptação os neurônios excitados têm seus pesos sinápticos ajustados. O ajuste feito é tal que a resposta do neurônio vencedor à aplicação de um padrão de entrada similar é melhorada.

2.4.1 O Algoritmo de Kohonen

O primeiro passo na execução do algoritmo de Redes Neurais de Kohonen é a inicialização, onde se definem a taxa de aprendizagem inicial $\eta(0)$, o raio de vizinhança inicial $\sigma(0)$, os pesos sinápticos iniciais dos neurônios $\underline{w}_j(0)$ e o número máximo de iterações N. Escolhem-se valores aleatórios para os vetores de pesos iniciais e recomenda-se a padronização dos dados no intervalo [0, 1].

No segundo passo, define-se o critério de parada, que pode ser um número máximo de iterações, um número de iterações sem a alteração dos valores da matriz peso, dentre outros.

O terceiro passo é o treinamento, que envolve as fases competitiva, cooperativa e adaptativa, onde cada padrão $\underline{x}$ deve ser apresentado à rede. O aprendizado pode ser seqüencial ou por lote. Na aprendizagem por lote a atualização dos pesos dos neurônios acontece ao final de cada iteração e na aprendizagem seqüencial a atualização ocorre após a apresentação de cada padrão.

Na fase competitiva, calculam-se as distâncias do padrão a todos os neurônios e verifica-se qual é o neurônio vencedor, ou seja, aquele cuja distância ao padrão seja a mínima. A distância é uma medida de dissimilaridade, ou seja, quanto menor for a distância, mais próximo o neurônio estará do padrão analisado. A distância Euclidiana é a medida de dissimilaridade mais comum. Nesta fase, também pode ser usada uma medida de similaridade. Neste caso, quanto maior a medida de similaridade, mais próximo o neurônio estará do padrão analisado. A

correlação é a medida de similaridade mais comumente utilizada. Outras medidas de similaridade e de distância são encontradas em Kohonen (1995).

Na fase cooperativa, localizam-se os vizinhos do neurônio vencedor. Na figura 2.9 são mostradas duas topologias: retangular e hexagonal (para 2D-SOM) e são identificados os vizinhos para raios de vizinhança iguais a zero, um e dois. Se o raio de vizinhança é zero, o neurônio não possui vizinhos e somente o neurônio vencedor é atualizado. Se o raio de vizinhança é um, na grade retangular cada neurônio tem oito vizinhos e na grade hexagonal cada neurônio tem seis vizinhos e assim sucessivamente.

Na fase adaptativa, atualizam-se os pesos sinápticos dos neurônios vizinhos ao neurônio vencedor segundo a equação 2.5, apresentada mais adiante. Nesta atualização leva-se em consideração a distância do vizinho até o neurônio vencedor e a atualização é mais intensa nos vizinhos mais próximos. Uma função de vizinhança, que varia com o tempo *n*, é utilizada neste sentido e deve satisfazer a duas exigências: ser simétrica em relação ao seu ponto máximo (que é atingido no neurônio vencedor) e decrescer monotonamente com o aumento da distância lateral. A função Gaussiana (equação 2.6) é a mais utilizada.

$$\underline{w}_j(n+1) = \underline{w}_j(n) + \eta(n) \cdot h_{j,i(x)}(n) \cdot (\underline{x} - \underline{w}_j(n)) \tag{2.5}$$

$$h_{j,i(x)}(n) = e^{\left(\frac{d_{j,i}^2}{2\sigma^2(n)}\right)} \tag{2.6}$$

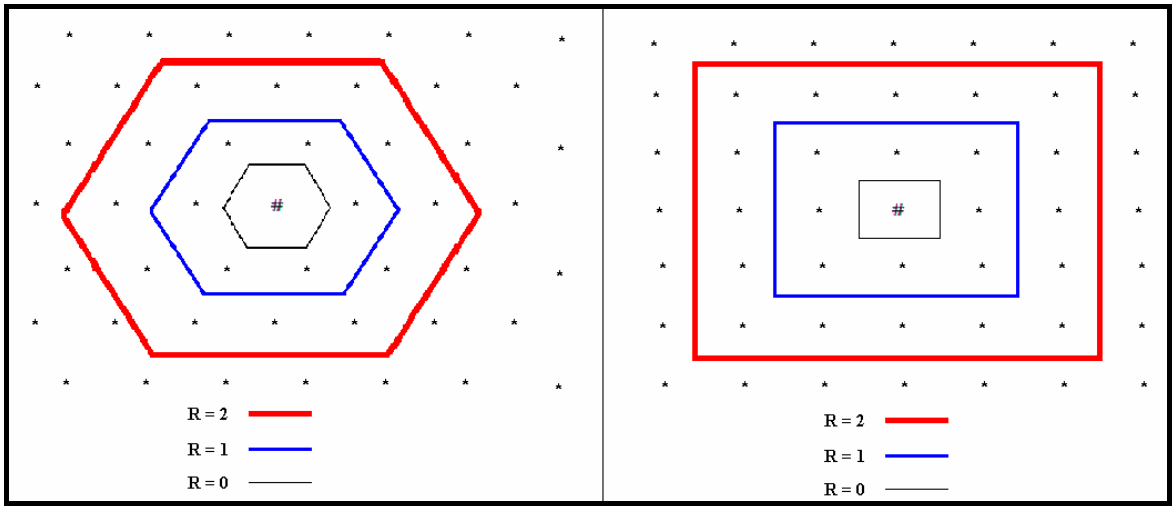

Figura 2.9 – Vizinhança para grades retangular e hexagonal com raios de vizinhança iguais a zero, um e dois (FAUSETT, 1994).

O quarto passo consiste em atualizar a taxa de aprendizagem (equação 2.7) e o raio de vizinhança (equação 2.8). Em seguida, retorna-se ao segundo passo até que um critério de parada seja satisfeito.

$$\eta(n) = \eta_0 e^{\frac{n}{\tau_2}} \tag{2.7}$$

$$\sigma(n) = \sigma_0 e^{\frac{n}{\tau_1}} \tag{2.8}$$

Segundo Siqueira (2005), podem ser utilizadas várias medidas de erro para determinar a qualidade de um mapa. O autor utiliza em seu trabalho o erro de quantização, que representa o erro médio correspondente à diferença entre os padrões e os pesos dos neurônios vencedores; o erro topológico, que representa o percentual de neurônios vencedores que não possuem o segundo vencedor em uma vizinhança de raio unitário centrada no neurônio vencedor e o erro médio quadrático.

Uma forma para determinar os agrupamentos utiliza a matriz de densidade. Nesta matriz cada elemento representa o número de padrões associados ao respectivo neurônio. Os neurônios com número reduzido de padrões associados a ele determinam as fronteiras entre os agrupamentos.

Existem diversas abordagens variantes das Redes Neurais de Kohonen. Os algoritmos, inspirados no original, modificam alguns aspectos como, por exemplo, critério de vizinhança, forma de escolha do neurônio vencedor, o uso de mapas hierárquicos, aceleração da aprendizagem, dentre outros (KOHONEN, 1995).

2.5 AGRUPAMENTO BASEADO EM FORMIGAS

2.5.1 Histórico

Sociedades de insetos sociais são sistemas distribuídos que apresentam uma organização social altamente estruturada, apesar da simplicidade dos seus indivíduos. Como resultado desta organização, colônias de formigas podem realizar tarefas complexas que, em alguns casos, excede a capacidade individual de uma única formiga. Na área de "algoritmos de formigas" estudam-se modelos inspirados na observação do comportamento de formigas reais e usam-se estes modelos como

fonte de inspiração para o desenvolvimento de novos algoritmos para a solução de problemas de otimização e de controle distribuído (DORIGO; STÜTZLE, 2004).

Entre os comportamentos dos insetos sociais, o mais amplamente reconhecido é a habilidade das formigas para trabalhar em grupo para desenvolver uma tarefa que não poderia ser executada por um único agente. Também visto em sociedade humana, esta habilidade de formigas é um resultado de efeitos cooperativos. O efeito cooperativo recorre ao fato de que o efeito de dois ou mais indivíduos ou partes coordenadas é mais alto do que o total dos efeitos individuais. Alguns pesquisadores alcançaram resultados promissores em mineração de dados usando uma colônia de formigas artificial. O número alto de indivíduos em colônias de formigas e a abordagem descentralizada para tarefas coordenadas (executadas de forma simultânea) significam que colônias de formigas mostram graus altos de paralelismo, auto-organização e tolerância a falhas. Estas características são desejadas em técnicas de otimização modernas (BORICZKA, 2009).

Muitos pesquisadores focalizaram sua atenção em uma classe nova de algoritmos, chamados de metaheurísticos. Segundo Dorigo e Stützle (2004), uma metaheurística é um conjunto de conceitos algorítmicos que podem ser usados para definir métodos heurísticos aplicáveis para um grande conjunto de diferentes problemas.

Um metaheurística, particularmente promissora, foi inspirada no comportamento de formigas reais. Começando com Sistema de Formigas, foram desenvolvidas e aplicadas várias abordagens algorítmicas baseadas nas mesmas idéias com sucesso considerável para uma variedade de problemas de otimização combinatória, acadêmicos e reais (DORIGO; STÜTZLE, 2004).

Otimização por colônia de formigas (*Ant Colony Optimization – ACO*) é uma metaheurística em que a colônia de formigas artificiais coopera para encontrar boas soluções para problemas de otimização discretos difíceis (DORIGO; STÜTZLE, 2004). Dorigo, Caro e Gambardella (1999) apresentam uma avaliação de trabalhos recentes em algoritmos de formiga para a otimização discreta e introduzem a metaheurística *ACO*. Dorigo e Blum (2005) apresentam uma pesquisa sobre resultados teóricos em otimização por colônia de formigas.

Segundo Boryczka (2009), muitos outros pesquisadores aplicaram o mecanismo de Otimização por Colônia de Formigas de Dorigo e Stützle a muitos

problemas de otimização combinatória e, então, o estenderam a uma classe inteira de problemas de otimização.

Socha e Dorigo (2008) apresentam uma extensão do *ACO* para domínios contínuos. No artigo, os autores mostram como o *ACO*, inicialmente desenvolvido para otimização combinatória, pode ser adaptado à otimização contínua sem qualquer mudança conceitual na sua estrutura. Os autores apresentam a idéia geral, a implementação e os resultados obtidos, os quais foram comparados com outros métodos de otimização contínua.

A fonte inspiradora do *ACO* é um comportamento observado em formigas reais. Ao procurar comida, inicialmente as formigas exploram aleatoriamente a área que cerca seu ninho. Quando uma formiga encontra uma fonte de comida, ela a leva para o ninho. No caminho, a formiga deposita um rastro de feromônio, cuja quantidade depende da quantidade e da qualidade da comida, que guia as outras formigas para a fonte de comida (SOCHA; DORIGO, 2008).

O Problema do Caixeiro Viajante (ou *Traveling Salesman Problem – TSP*) é um problema muito estudado na literatura. O *TSP* também tem um papel importante na pesquisa de *ACO*: o primeiro algoritmo de *ACO*, chamado Sistema de Formiga (*Ant System* - *AS*) foi testado primeiro no *TSP* (DORIGO; STÜTZLE, 2004).

Segundo Dorigo, Maniezzo e Colorni (1996), na escolha de um trajeto, uma formiga é influenciada pela intensidade dos rastros de feromônio. Um nível mais alto de feromônio dá para uma formiga um estímulo mais forte e assim uma probabilidade mais alta para escolhê-lo. O resultado é que uma formiga encontrará um rastro mais forte em caminhos mais curtos. Como conseqüência, o número de formigas que seguem estes caminhos será mais alto. Isto fará com que a quantidade de feromônio no caminho mais curto cresça mais rápido do que no mais longo e, então, a probabilidade com que qualquer formiga escolhe um caminho para seguir é rapidamente tendenciada para o mais curto. O resultado final é que muito depressa todas as formigas escolherão o caminho mais curto.

Sistema de Colônia de Formigas (*Ant Colony System – ACS*), descrito por Dorigo e Gambardella (1997), difere do *AS* em três pontos principais. Primeiro, explora mais fortemente a experiência de busca acumulada pelas formigas. Segundo, evaporação de feromônio e depósito de feromônio somente ocorrerão nos arcos que pertencem ao melhor caminho até o momento. Terceiro, cada vez que

uma formiga usa um arco, remove algum feromônio do mesmo, aumentando a exploração de caminhos alternativos (DORIGO; STÜTZLE, 2004).

O estudo de colônias de formigas tem oferecido notável contribuição, não só na otimização combinatória, mas também oferecendo idéias novas para técnicas de agrupamento (Boriczka, 2009).

O Agrupamento baseado em Formigas foi proposto inicialmente por Deneubourg *et al.* (1991, *apud* Handl, Knowles e Dorigo, 2006). Em contraste com o *ACO*, nenhum feromônio artificial é usado, sendo que o próprio ambiente serve como variável de estimergia (DORIGO; BONABEAU; THERAULAZ, 2000).

2.5.2 As Operações de Carregar e Descarregar Padrões

No Agrupamento baseado em Formigas proposto por Deneubourg et al. (1991, *apud* Handl, Knowles e Dorigo, 2006), as formigas foram representadas como agentes simples que se moviam aleatoriamente em uma grade quadrada. Os padrões foram dispersos dentro desta grade e poderiam ser carregados, transportados e descarregados pelos agentes (formigas). Estas operações são baseadas na similaridade e na densidade dos padrões distribuídos dentro da vizinhança local dos agentes, padrões isolados ou cercados por dissimilares são mais prováveis de serem carregados e então descarregados numa vizinhança de similares.

As decisões de carregar e descarregar padrões são tomadas pelas probabilidades P_{pick} e P_{drop} dadas pelas equações 2.9 e 2.10, a seguir, respectivamente.

$$P_{pick} = \left(\frac{k_p}{k_p + f(i)} \right)^2 \tag{2.9}$$

$$P_{drop} = \left(\frac{f(i)}{k_d + f(i)} \right)^2 \tag{2.10}$$

Nestas equações, *f(i)* é uma estimativa da fração de padrões localizados na vizinhança que são semelhantes ao padrão atual da formiga e k_p e k_d são constantes reais. No trabalho de Deneubourg *et al.* (1991, *apud* Handl, Knowles e Dorigo,

2006), os autores usaram k_p = 0,1 e k_d = 0,.3. Neste trabalho, os autores obtiveram a estimativa *f*, através de uma memória de curto prazo de cada formiga, onde o conteúdo da última célula da grade analisada é armazenado. Esta escolha da função de vizinhança *f(i)* foi essencialmente motivada pela sua facilidade de realização por robôs simples.

Lumer e Faieta (1994, *apud* Handl, Knowles e Dorigo, 2006) introduziram um número de modificações ao modelo que permitiu a manipulação de dados numéricos e melhorou a qualidade da solução e o tempo da convergência do algoritmo. A idéia era definir uma medida de similaridade ou dissimilaridade entre os padrões, já que no algoritmo proposto inicialmente, os objetos eram similares se os objetos fossem idênticos e dissimilares se os objetos não fossem idênticos. No referido trabalho aparece pela primeira vez o mapeamento topográfico.

Segundo Vizine *et al.* (2005), a idéia geral deste algoritmo é ter dados semelhantes no espaço *n*-dimensional original em regiões vizinhas da grade, ou seja, dados que são vizinhos na grade indicam padrões semelhantes no espaço original.

No trabalho de Lumer e Faieta (1994, *apud* Handl, Knowles e Dorigo, 2006), a decisão de carregar padrões é baseada na probabilidade P_{pick} dada pela equação 2.9 anterior e a decisão de descarregar padrões é baseada na probabilidade P_{drop} dada pela equação 2.11 a seguir, onde *f(i)* é dada pela equação 2.12.

$$P_{drop} = \begin{cases} 2f \quad , \quad se \qquad f(i) < k_d \\ 1 \quad , \quad se \qquad f(i) \geq k_d \end{cases} \tag{2.11}$$

$$f(i) = \max\left\{0, \frac{1}{\sigma^2} \sum_{j \in L}\left[1 - \frac{d(i,j)}{\alpha}\right]\right\} \tag{2.12}$$

Na equação 2.12, *d*(*i*, *j*) é uma função de dissimilaridade entre padrões *i* e *j* pertencentes ao intervalo [0, 1]; *α* é um parâmetro escalar dependente dos dados (padrões) e pertencente ao intervalo [0, 1]; *L* é a vizinhança local de tamanho igual a σ^2, onde *σ* é o raio de percepção (ou vizinhança). Os autores usaram em seu trabalho k_p = 0,1, k_d = 0,15 e *α* = 0,5.

Os algoritmos de Agrupamento baseados em Formigas estão principalmente baseados nas versões propostas por Deneubourg *et al.* (1991, *apud* Handl, Knowles

e Dorigo, 2006) e Lumer e Faieta (1994, *apud* Handl, Knowles e Dorigo, 2006). Várias modificações foram introduzidas para melhorar a qualidade do agrupamento e, em particular, a separação espacial entre os grupos na grade (BORICZKA, 2009).

Mudanças que melhoram a separação espacial dos grupos e permitem que o algoritmo seja mais robusto foram introduzidas por Handl, Knowles e Dorigo (2006). Uma delas é a restrição na função *f(i)* dada pela equação 2.13, a seguir, que serve para penalizar dissimilaridades elevadas.

$$f^*(i) = \begin{cases} \frac{1}{\sigma^2} \sum_{j \in L} \left[1 - \frac{d(i,j)}{\alpha} \right] & , \quad se \quad \forall j \quad \left(1 - \frac{d(i,j)}{\alpha} \right) > 0 \\ 0 & , \quad caso \qquad contrário \end{cases} \tag{2.13}$$

Segundo Vizine *et al.* (2005), uma dificuldade na aplicação do algoritmo de Agrupamento por Formigas em problemas complexos é que, na maioria dos casos, eles geram um número de grupos muito maior que o real. Além disso, estes algoritmos normalmente não estabilizam em uma solução de agrupamento, ou seja, eles constantemente constroem e desconstroem grupos durante o processo. Para superar estas dificuldades e melhorar a qualidade dos resultados, os autores propuseram um Algoritmo de Agrupamento por Formigas Adaptável (*Adaptative Ant Clustering Algoritthm* - A^2CA). Uma modificação incluída nesta abordagem é um programa de resfriamento para o parâmetro que controla a probabilidade de formigas apanharem objetos da grade.

2.5.3 Parâmetros da Função de Vizinhança

A separação espacial dos grupos na grade é crucial para que grupos individuais sejam bem definidos, permitindo a sua recuperação automática. A proximidade espacial, quando ocorrer, pode indicar a formação prematura do agrupamento (HANDL; KNOWLES; DORIGO, 2006).

A definição dos parâmetros da função de vizinhança é um fator decisivo na qualidade do agrupamento. No caso do raio de percepção σ, é mais atrativo empregar vizinhanças maiores para melhorar a qualidade do agrupamento e da distribuição na grade. Porém, este procedimento é mais caro computacionalmente

(porque o número das células a serem consideradas para cada ação cresce quadraticamente com o raio), e ainda inibe a formação rápida dos grupos durante a fase de distribuição inicial. Um raio da percepção que aumenta gradualmente com o tempo acelera a dissolução de grupos pequenos preliminares (HANDL; KNOWLES; DORIGO, 2006). Um raio de percepção progressivo também foi utilizado por Vizine *et al.* (2005).

Além disso, depois de uma fase inicial de agrupamento, Handl, Knowles e Dorigo (2006) substituíram o parâmetro escalar $\frac{1}{\sigma^2}$ por $\frac{1}{N_{occ}}$ na equação 2.13, onde N_{occ} é o número de células da grade ocupadas, observadas dentro da vizinhança local. Assim, somente a semelhança e não a densidade foi levada em conta. Boryczka (2009), em seu algoritmo *ACAM* (do inglês "*Ant-based clustering algorithm*", que significa "Algoritmo de Agrupamento Baseado em Formigas Modificado"), propôs a substituição do escalar $\frac{1}{\sigma^2}$ na equação 2.13 pelo escalar $\frac{\sigma_0^2}{\sigma^2}$, onde σ_0 é o raio de percepção inicial.

Segundo Handl, Knowles e Dorigo (2006), α determina a porcentagem de padrões na grade classificados como semelhantes. A escolha de um valor muito pequeno para α, impede a formação de grupos na grade; por outro lado, a escolha de um valor muito grande para α, resulta na fusão de grupos.

Fixar parâmetro α não é simples e a sua escolha é altamente dependente da estrutura do conjunto de dados. Um valor inadequado é refletido por uma excessiva ou extremamente baixa atividade na grade. A quantidade de atividade é refletida pela freqüência de operações com sucesso da formiga em carregar e descarregar. Com base nestas análises, Handl, Knowles e Dorigo (2006) propuseram uma adaptação automática de α. Já Boryczka (2009) propôs um novo esquema de adaptação para o valor de *α*.

Tan, Ting e Teng (2007) examinam o parâmetro escalar de dissimilaridade em abordagens de Colônia de Formigas para agrupamento de dados. Os autores mostram que não há necessidade de usar uma adaptação automática de α. Os mesmos propõem um método para calcular um α fixo para cada base de dados. O valor α é calculado independentemente do processo de agrupamento.

Para medir a similaridade entre os padrões, diferentes métricas são utilizadas. Handl, Knowles e Dorigo (2006) utilizam distância Euclidiana para dados sintéticos e Co-seno para dados reais. Boryczka (2009) testou diferentes medidas de dissimilaridade: Euclidiana, Co-seno e medidas de Gower.

2.5.4 A Memória de Curto Prazo

O processo de agrupamento também pode ser acelerado significativamente pelo uso de uma memória de curto prazo introduzida por Lumer e Faieta (1994, *apud* Handl, Knowles e Dorigo, 2006). Cada agente recorda os últimos padrões carregados e as respectivas posições onde foi descarregado. Quando um padrão novo é carregado, a posição de "*best matching*" (melhor combinação) memorizada será usada para indicar o sentido da direção aleatória do agente. O "*best matching*" é a posição de mínima dissimilaridade.

Na proposta de Handl, Knowles e Dorigo (2006), esta memória de curto prazo foi estendida. Os padrões armazenados na memória podem já ter sido removidos da posição registrada. A fim de determinar mais robustamente o sentido da direção, os autores permitem que cada agente "olhe adiante". Um agente que carrega um padrão *i* usa sua memória para examinar todas as posições registradas, uma após a outra, usando a função de vizinhança *f* (*i*). O "*best match*" é a célula da grade para qual a função da vizinhança rende o valor mais elevado. Entretanto, o salto para o "*best match*" só é feito com alguma probabilidade, dependente da qualidade do "*match*" (combinação).

Inspirados em alguns aspectos do fenômeno da trofalaxis (troca de comida líquida entre insetos), Oca, Garrido e Aguirre (2005) estudaram duas estratégias de comunicação direta entre agentes: compartilhamento de memória e de mapas ambientais. Como se deseja que os agentes troquem informação sobre a distribuição espacial dos dados no ambiente, é permitido a eles enriquecer ou atualizar suas próprias representações do ambiente. O objetivo é permitir que os agentes escolham a melhor localidade onde depositar um dado (padrão) e, portanto, criar melhores agrupamentos. Os autores também mostram que os resultados dependem da densidade de agentes no ambiente e da utilidade da informação trocada.

2.5.5 A Inclusão do Feromônio

Sherafat *et al.* (2004, *apud* Vizine *et al.*, 2005) introduziram uma variável local *Φ(i)* associada a cada posição *i* na grade, tal que a quantidade de feromônio naquela posição se torna uma função da presença ou ausência de um objeto (padrão) em *i*. Os agentes artificiais no algoritmo de Agrupamento por Formigas proposto acrescentarão algum feromônio aos objetos carregados por eles e este feromônio será transferido à grade quando os objetos são depositados. Para reduzir o número de parâmetros definidos pelo usuário e melhorar o desempenho do algoritmo, Vizine *et al.* (2005) propuseram a substituição das funções P_{pick} e P_{drop} dadas por Sherafat *et al.*, pelas equações 2.14 e 2.15, respectivamente. Durante cada iteração, o feromônio artificial *Φ(i)* em cada célula da grade *i*, evapora a uma taxa fixa, dada pela equação 2.16.

$$P_{pick(i)} = \frac{1}{f(i)\phi(i)}\left(\frac{k_p}{k_p + f(i)}\right)^2 \tag{2.14}$$

$$P_{drop(i)} = f(i)\phi(i)\left(\frac{f(i)}{k_d + f(i)}\right)^2 \tag{2.15}$$

$$\phi(i) \longleftarrow \phi(i).0,99 \tag{2.16}$$

Vizine *et al.* (2005), utilizaram uma função de feromônio adicionada à grade como um modo para promover um reforço para derrubar objetos em regiões mais densas da grade.

2.5.6 Outras Abordagens

Yang e Kamel (2006) apresentam uma abordagem de múltiplas colônias de formigas para agrupamento que consiste em algumas colônias de formigas paralelas e independentes e um agente “formiga rainha”. Cada processo de colônia de formigas gera um resultado de agrupamento com um algoritmo de Agrupamento baseado em Formigas. Estes resultados são enviados ao agente “formiga rainha” e combinados para calcular uma matriz de similaridade nova. A nova matriz é devolvida para cada processo de colônia de formigas para reagrupar os dados

usando a informação nova. Resultados experimentais mostraram que o desempenho médio dos algoritmos de múltiplas colônias de formigas ultrapassou o algoritmo de Agrupamento baseado em Formiga simples e o algoritmo *k*-médias.

Existem ainda outras vertentes de agrupamento por formigas. Segundo Handl e Meyer (2007), existem dois tipos principais de agrupamento baseado em formigas. O primeiro grupo imita diretamente o comportamento observado no agrupamento de colônias de formigas reais. Já o segundo grupo é indiretamente inspirado pela natureza, pois a tarefa de agrupamento é reformulada como uma tarefa de otimização e geralmente heurísticas de otimização baseada em formigas são utilizadas para encontrar agrupamentos bons ou próximos do ótimo. Este artigo traz uma revisão de ambas as abordagens.

Em Azzag *et al.* (2007), um modelo novo de agrupamento (denominado *AntTree*) é apresentado. Este modelo é baseado no comportamento de formigas reais e pode ser usado para construir um agrupamento hierárquico estruturado em árvores. O algoritmo proposto obteve resultados competitivos quando comparado a outros métodos. *AntTree* foi aplicado a três situações reais.

Kuo *et al.* (2005) propuseram um novo método de agrupamento denominado *Ant K-means* (*AK*). O algoritmo *AK* modificou o k-médias localizando os objetos em um grupo com uma probabilidade que é atualizada pelo feromônio. Os resultados mostraram que o método proposto foi melhor que dois outros métodos.

Ghosh *et al.* (2008) apresentam um algoritmo novo (chamado APC) para agrupamento de dados baseado na propriedade de agregação de feromônio encontrado em formigas. Cada formiga representa um dado. Cada formiga é colocada em um local e as formigas se movem para encontrar pontos com densidade de feromônio mais alta. O movimento de uma formiga é baseado na quantidade de feromônio depositado em diferentes pontos do espaço de busca. Quanto maior o feromônio depositado, maior é a agregação de formigas e isto conduz à formação de grupos de dados homogêneos. O algoritmo ligação média foi aplicado após a formação dos grupos para obter o número de grupos desejado. O algoritmo proposto foi avaliado utilizando-se várias bases de dados conhecidas usando medidas de validade de agrupamento diferentes e foi comparado com duas técnicas de agrupamento populares. Resultados experimentais justificam a potencialidade do algoritmo proposto.

O algoritmo proposto por Fernandes *et al.* (2008) (denominado *KohonAnts*) é um algoritmo de otimização por colônia de formigas inspirado nos conceitos de mapas auto-organizáveis e algoritmos por formigas. É baseado em várias novas idéias. Neste algoritmo cada formiga representa um dado. Formigas deslocam-se na grade deixando “feromônios vetoriais”. A grade é preenchida inicialmente com vetores feromônios aleatórios (com a mesma dimensão dos dados) e cada vez que uma formiga cai em uma célula, ela muda o feromônio seguindo um método semelhante ao utilizado em mapas auto-organizáveis de Kohonen, fazendo com que o feromônio da célula fique mais perto do dado armazenado na própria formiga. Como as formigas se deslocam na grade, a posição e o feromônio da formiga têm conteúdo co-adaptado, para que formigas com dados semelhantes estejam próximas na grade e que a grade, por sua vez, contenha vetores semelhantes aos armazenados na formiga sobre ela. A grade pode então ser usada para classificar, da mesma forma que os mapas auto-organizáveis de Kohonen, porém, com melhores resultados. Além disso, as formigas podem ser usadas para identificar visualmente a posição dos grupos.

2.5.7 O Algoritmo Básico proposto por Deneubourg *et al.* (1991, *apud* Handl, Knowles e Dorigo, 2006)

Numa fase inicial, todos os padrões são aleatoriamente espalhados na grade. Depois, cada formiga escolhe aleatoriamente um padrão para carregar e é colocada em uma posição aleatória na grade.

Na próxima fase, chamada de fase de distribuição, em um laço (*loop*) simples, cada formiga é selecionada aleatoriamente. Esta formiga se desloca na grade executando um passo de comprimento *L*, numa direção determinada aleatoriamente. Segundo Handl, Knowles e Dorigo (2006), o uso de um tamanho de passo grande acelera o processo de agrupamento. A formiga então decide, probabilisticamente, se descarrega seu padrão nesta posição.

Se a decisão de descarregar o padrão for negativa, escolhe-se aleatoriamente outra formiga e recomeça-se o processo. No caso de decisão positiva, a formiga descarrega o padrão em sua posição atual na grade, se esta estiver livre. Se esta célula da grade estiver ocupada por outro padrão, o mesmo

deve ser descarregado numa célula imediatamente vizinha desta, que esteja livre, por meio de uma procura aleatória.

A formiga procura, então, por um novo padrão para carregar. Dentre os padrões livres na grade, ou seja, dentre os padrões que não estão sendo carregados por nenhuma formiga, a formiga seleciona aleatoriamente um, vai para a sua posição na grade, faz a avaliação da função de vizinhança e decide probabilisticamente se carrega este padrão. Este processo de escolha de um padrão livre na grade é executado até que a formiga encontre um padrão que deva ser carregado.

Só então esta fase é reiniciada, escolhendo-se outra formiga até que um critério de parada seja satisfeito.

2.5.8 Recuperação do Agrupamento

O processo inicia com cada padrão formando um grupo. Depois de calcular as distâncias entre todos os grupos, deve-se fundir (ligar) os dois grupos com menor distância. Os tipos de ligações mais comuns são: Ligação Simples, Ligação Completa, Ligação Médias e Método de Ward (JOHNSON; WICHERN, 1998). As distâncias entre os grupos são definidas em termos de sua distância na grade. Cada padrão é agora composto de apenas dois atributos, que o posicionam na grade bidimensional. A distância entre cada dois padrões é então a distância Euclidiana entre dois pontos da grade. Este processo se repete até que um critério de parada seja satisfeito.

Quando padrões em torno das bordas dos grupos estão isolados, Handl, Knowles e Dorigo (2006) introduziram um peso que incentiva a fusão destes padrões com os grupos.

2.6 AGRUPAMENTOS EM SÉRIES TEMPORAIS

Segundo Liao (2005), a maioria das análises de agrupamentos foram realizadas em dados estáticos. Os dados são chamados estáticos se todas as suas características não mudam ou mudam insignificantemente com o tempo. Diferentemente de dados estáticos, atributos de séries temporais compreendem valores que mudaram com o tempo. Séries temporais são de interesse devido à sua

presença em diversas áreas da ciência, economia, finanças, economia, saúde, administração pública e engenharia (onde se enquadra nosso principal objeto de estudo, a instrumentação geotécnica-estrutural de uma grande barragem). Obras dedicadas à análise de agrupamento de séries temporais são relativamente escassas em comparação com dados estáticos, no entanto, parece haver uma tendência em aumentar esta atividade.

Segundo Liao (2005), as primeiras abordagens trabalharam diretamente com dados de séries temporais e a principal modificação foi a substituição da medida de distância ou similaridade para dados estáticos por uma apropriada para séries temporais. As abordagens posteriores converteram os dados de séries temporais em um vetor de características de dimensão mais baixa ou vários parâmetros modelo e, então, aplicaram um algoritmo de agrupamento convencional. Métodos de Particionamento, hierárquicos e baseados em modelos, têm sido utilizados diretamente ou propostos para o agrupamento de séries temporais.

Steinbach *et al.* (2002) aplicaram Mineração de dados à descoberta de índices climáticos do oceano (*Ocean Climate Indices – OCI)*. Estes índices são séries temporais que resumem o comportamento de áreas selecionadas dos oceanos na Terra e constituem uma ferramenta importante para predizer o efeito dos oceanos no clima da terra. Foi aplicado o algoritmo de agrupamento vizinho mais próximo para agrupar séries temporais de pressão e temperatura associadas a pontos do oceano, agrupando regiões do oceano com comportamento relativamente homogêneo. Os centros destes grupos são séries temporais que resumem o comportamento destas áreas do oceano e representam *OCI*'s potenciais.

Diniz *et al.* (2003) utilizam análise de agrupamento para a determinação de regiões homogêneas de temperaturas máxima e mínima para o Estado do Rio Grande do Sul. Os autores afirmam que a determinação de regiões homogêneas de variáveis meteorológicas tem sido utilizada em várias pesquisas climatológicas. Esta determinação ajuda no zoneamento agroclimático e serve de subsídio ao planejamento agrícola das regiões produtoras do estado. Foram testados quatro métodos hierárquicos de agrupamento: Ligação Simples, Ligação Completa, Método do centróide e Método Ward. O método adotado foi o de Ligação Completa. Foram obtidas quatro regiões representativas. A climatologia das regiões foi feita pelo cálculo da média das séries temporais das estações contidas em cada região homogênea. A técnica e o número de grupos obtidos foram satisfatórios no processo

de identificação e separação das regiões homogêneas de temperaturas máxima e mínima, representando as condições fisiográficas do Estado do Rio Grande do Sul.

Liao (2005) faz uma revisão sobre agrupamento de séries temporais mostrando algoritmos utilizados para esta tarefa, medidas para calcular a distância ou similaridade entre duas séries temporais e critérios para determinar a qualidade do agrupamento é obtido. Métodos de agrupamento hierárquico (como o Método Ward), mapas auto-organizáveis de Kohonen, Distância Euclidiana e medida de similaridade (*Sim*), são apresentados pelo autor e serão utilizados neste trabalho.

2.7 AVALIAÇÃO DO AGRUPAMENTO

Na avaliação de grupos, diferentes aspectos podem ser observados: determinação da tendência de agrupamento de um conjunto de dados, comparação dos resultados de uma análise de grupos com resultados externamente conhecidos, avaliação de quão bem os resultados de uma análise de grupos se ajustam aos dados sem referência a informação externa, comparação dos resultados de dois diferentes conjuntos de análise de grupos para determinar qual deles é melhor ou, ainda, determinação do número correto de grupos (TAN; STEINBACH; KUMAR, 2005).

Segundo Tan, Steinbach e Kumar (2005), as medidas numéricas aplicadas para julgar vários aspectos de avaliação de grupos são classificadas em três tipos: os índices externos são usados para medir até que ponto rótulos de grupos correspondem a rótulos de classes externamente fornecidos; os índices internos são usados para medir quão boa é a estrutura de agrupamento sem relação com informação externa e os índices relativos são usados para comparar dois grupos ou agrupamentos diferentes.

Boryczka (2009) utilizou em seu trabalho dois índices internos (a Variância Intra-Grupos e o Índice *Dunn*) e dois índices externos (a medida *F* e o Índice Aleatório). Estas medidas são descritas, a seguir, e utilizadas também neste trabalho.

A Variância é a soma dos quadrados dos desvios entre todos os padrões e os respectivos centróides do grupo à que pertencem, dividido pelo número total de padrões. A Variância deve ser minimizada.

O Índice Dunn (*D*) determina a razão mínima entre a distância intra-grupo e o diâmetro do grupo para um determinado agrupamento. *D* é definido pela equação 2.17, onde *C* é o conjunto de todos os grupos, *diam(c)* é o diâmetro de um grupo *c* e $\delta(\mu_c, \mu_d)$ é a distância entre os centróides dos grupos *c* e *d*. O diâmetro de um grupo *c* é calculado como a máxima distância intra-grupo. *D* deve ser maximizado.

$$D = \min_{c,d \in C} \left[\frac{\delta(\mu_c, \mu_d)}{\max_{e \in C} \left[diam(e) \right]} \right] \tag{2.17}$$

A medida *F* usa a idéia de precisão e memória da recuperação da informação. Cada classe *i* é um conjunto de n_i padrões desejados; cada grupo *j* (gerado pelo algoritmo) é um conjunto de n_j padrões; n_{ij} é o número de padrões da classe *i* pertencentes ao grupo *j*. Para cada classe *i* e grupo *j*, a precisão *p* e a memória *r* são definidas como $p(i,j) = \frac{n_{ij}}{n_j}$ e $r(i,j) = \frac{n_{ij}}{n_i}$, respectivamente. O valor da medida *F* é dado pela equação 2.18.

$$F = \sum_i \frac{n_i}{n} \max_j \{F(i,j)\} \tag{2.18}$$

onde:

$$F(i,j) = \frac{(b^2+1).p(i,j).r(i,j)}{b^2.p(i,j) + r(i,j)}$$

O valor de *b* deve ser “1” para proporcionar pesos iguais para a precisão *p* e a recordação *r*. Na equação 2.18, *n* é o tamanho do conjunto de dados. *F* é limitada ao intervalo [0, 1] e deve ser maximizada.

O Índice Aleatório (*R*) é dado pela equação 2.19, onde *a*, *b*, *c* e *d* são calculados para todos os possíveis pares de padrões *i* e *j* e seus respectivos grupos *U* (classificação correta - $c_U(i)$ e $c_U(j)$) e *V* (solução gerada pelo algoritmo de agrupamento – $c_V(i)$ e $c_V(j)$). *R* está limitado no intervalo [0, 1] e deve ser maximizado.

$$R = \frac{a+d}{a+b+c+d} \tag{2.19}$$

onde:

$$a = \left| \left\{ i, j \mid c_U(i) = c_U(i) \hat{} \, c_V(i) = c_V(i) \right\} \right|$$
$$b = \left| \left\{ i, j \mid c_U(i) = c_U(i) \hat{} \, c_V(i) \neq c_V(i) \right\} \right|$$
$$c = \left| \left\{ i, j \mid c_U(i) \neq c_U(i) \hat{} \, c_V(i) = c_V(i) \right\} \right|$$
$$d = \left| \left\{ i, j \mid c_U(i) \neq c_U(i) \hat{} \, c_V(i) \neq c_V(i) \right\} \right|$$

Para avaliar os resultados de agrupamentos em séries temporais, Keogh *et al.* (2006) utilizaram a medida de similaridade (*Sim*), que é descrita a seguir e, também, utilizada neste trabalho. Considerando-se a classe *C* (rótulo) e o grupo *C'* (resultado de uma abordagem particular), calcula-se a semelhança entre eles usando a fórmula da equação 2.20.

$$Sim(C, C') = \frac{\left(\sum_i \max_j Sim(C_i, C_j) \right)}{K} \tag{2.20}$$

onde:

$$Sim(C_i, C'_j) = \frac{\left| C_i \cap C'_j \right|}{\left| C_i \right| + \left| C'_j \right|}$$

Esta medida de semelhança será “0”, se o dois grupos forem completamente diferentes e “1”, se eles forem idênticos. Esta medida é descrita e utilizada em Gavrilov *et al.* (2000).

3 MATERIAIS E MÉTODOS

Com o intuito de melhor compreender as técnicas de Mineração de Dados (agrupamento de padrões), do processo *KDD*, aqui abordadas (Estatística Multivariada; Redes Neurais de Kohonen e Colônia de Formigas), as mesmas foram, inicialmente, aplicadas a bases de dados reais e de séries temporais. Além disso, estas técnicas foram aplicadas a uma base de dados real e inédita, a base de dados de instrumentação geotécnica-estrutural da Itaipu.

Desta forma, inicia-se este capítulo 3 com a apresentação das bases de dados (reais, séries temporais e de Itaipu) e, em seguida, cada etapa do processo *KDD* aplicada aos dados de Itaipu (Seleção dos Dados; Pré-processamento e Formatação dos Dados) é analisada. Na etapa de Mineração dos Dados (agrupamento de padrões), as referidas técnicas: Análise Estatística Multivariada; Redes Neurais de Kohonen e Colônia de Formigas, apresentadas no capítulo 2 são analisadas, definindo a melhor forma de aplicá-las às bases e, finalmente, é proposto o Algoritmo para Agrupamento baseado em Colônia de Formigas, com o objetivo de melhorar o seu desempenho. Os resultados numéricos são apresentados no capítulo 4.

3.1 BASES DE DADOS ABORDADAS

Foram abordadas seis bases de dados reais. As bases de dados IRIS; WINE e PIMA Indians Diabetes, são bases de dados reais e públicas, disponíveis em http://mlearn.ics.uci.edu/databases, e foram aqui denominadas “bases de dados reais”. As bases de dados GUN e LIGHTNING-2, também são bases de dados reais e públicas, disponíveis em Keogh *et al.* (2006), e foram denominadas “bases de dados de séries temporais”. A base de dados de instrumentação geotécnica-estrutural da barragem de Itaipu, obtida junto a equipe de engenheiros da Itaipu, é uma base de dados especialmente selecionada, pré-processada e formatada para este trabalho.

3.1.1 Bases de Dados Reais e de Séries Temporais

As bases de dados reais e de séries temporais abordadas são bases já utilizadas para a tarefa de agrupamento. As bases de dados reais foram escolhidas

pois já foram anteriormente utilizadas para agrupamento por Formigas. As bases de dados de séries temporais foram escolhidas pois, assim como a base de dados inédita deste trabalho, são séries temporais. Estas bases estão detalhadamente descritas no Anexo 1 deste trabalho, sendo que o quadro 3.1, a seguir, mostra o número de padrões, o número de atributos e a quantidade de grupos para cada uma destas bases de dados.

Nas bases de dados de séries temporais, os métodos de agrupamento foram aplicados diretamente, sem a aplicação de um método de pré-processamento dos dados visando o agrupamento dos dados especificamente para séries temporais.

	Base de Dados	**N° de Padrões**	**N° de atributos**	**N° de grupos**
Reais	IRIS	150	4	3
	WINE	178	13	3
	PIMA Indians Diabetes	768	8	2
Séries Temporais	GUN	200	150	2
	LIGHTNING-2	121	637	2

Quadro 3.1 – Bases de dados utilizados para avaliação dos algoritmos.

Conforme já comentado, estas bases de dados reais e de séries temporais foram utilizadas com o intuito de melhor entender todos os algoritmos analisados e poder comparar os resultados obtidos com os publicados na literatura. Com este embasamento, as propostas ao Algoritmo de Agrupamento baseado em Formigas puderam ser apresentadas, assim como a sua posterior aplicação aos dados de instrumentação da barragem de Itaipu, principal objetivo deste trabalho.

3.1.2 Base de Dados de Instrumentação Geotécnica-Estrutural da Itaipu

A barragem de Itaipu possui 7.919 m de extensão e altura máxima de 196 m, dimensões que transformaram esta obra em referência nos estudos de concreto e na segurança de barragens. A barragem de Itaipu é composta por dois trechos de barragens de terra, um trecho de barragem de enrocamento e trechos de concreto, compondo as estruturas mais altas do conjunto.

Em toda sua extensão, para acompanhar o desempenho das estruturas de concreto e fundação, são encontrados 2.218 instrumentos (1.362 no concreto e 856 nas fundações e aterros) sendo que destes, 270 estão automatizados; além disso,

existem 5.239 drenos (949 no concreto e 4.290 nas fundações). Todas estas leituras ocorrem em diferentes freqüências, podendo ser, por exemplo, diária, semanal, quinzenal, mensal, dependendo do tipo de instrumento, e vêm sendo armazenadas há mais de 30 anos. Detalhes adicionais sobre a barragem de Itaipu encontram-se no Anexo 2 deste trabalho.

É importante selecionar as informações que melhor “entendam” o comportamento da barragem, permitindo a previsão e a resolução de eventuais problemas que possam ocorrer. Neste trabalho, os instrumentos selecionados para análise foram os extensômetros localizados no trecho F da referida barragem, conforme detalhamento na seção 3.2 a seguir.

3.2 SELEÇÃO DOS DADOS

A seleção dos dados é a 1ª etapa do processo *KDD*, conforme figura 2.7, já apresentada. Embora todos os trechos da barragem sejam instrumentados e monitorados, um trecho da barragem chamado de Barragem Principal (denominado trecho F e identificado pelo número “5” na figura 1 do anexo 2) merece destaque e um estudo mais aprofundado. No trecho F encontram-se as turbinas para a geração da energia elétrica, além de ser o trecho de maior altura em coluna de água e o mais instrumentado. Este trecho é constituído de vários blocos, sendo que cada um deles possui instrumentos que fornecem dados a respeito de seu comportamento físico, tanto na estrutura de concreto como na sua fundação. Este trabalho foi desenvolvido baseado nos dados obtidos neste trecho da barragem.

No trecho F encontram-se extensômetros, piezômetros, medidores tri-ortogonais, medidores de nível de água e medidores de vazão. Destes, foram selecionados para análise os extensômetros que são do tipo múltiplo de hastes, instalados em furos de sondagem. Este tipo de instrumento é considerado um dos mais importantes, pois são os responsáveis pelas medições de recalques (deslocamentos verticais), que consistem em uma das observações mais importantes na supervisão do comportamento da estrutura da barragem.

No trecho F estão localizados 30 extensômetros. Cada um desses 30 extensômetros possui uma, duas ou três hastes, totalizando 72 medidas de deslocamento. Estas medidas serão identificadas da seguinte forma: por exemplo, equip4_1, significa haste 1 do extensômetro 4.

Os dados utilizados para o desenvolvimento deste trabalho são mensais e datados de janeiro/1995 a dezembro/2004, totalizando 120 leituras. O período foi assim determinado por sugestão da equipe de engenheiros da Itaipu, pois é bem posterior à construção da barragem e anterior à implantação do sistema de aquisição automática de dados, já que durante a fase de instalação deste sistema, alguns instrumentos ficaram sem leituras manuais e, além disso, os 11 instrumentos automatizados (totalizando 24 hastes) tiveram modificações que podem ter influenciado as leituras posteriores. Houve a troca da cabeça do instrumento por uma 70 cm mais longa. Desta forma, as referidas 120 leituras ficaram isentas a estas irregularidades.

3.3 PRÉ-PROCESSAMENTO E FORMATAÇÃO DOS DADOS

O pré-processamento dos dados é a 2ª etapa do processo *KDD* (figura 2.7). Os dados foram disponibilizados pela Itaipu em arquivos do tipo texto (**.txt*). Estes arquivos não puderam ser convertidos diretamente para planilhas, pois o número de linhas excedia o seu limite. Assim, primeiramente, dentre estes dados, foram selecionados somente aqueles que pertenciam ao período delimitado para este trabalho e, então, os dados foram convertidos para as planilhas. Destes dados foram extraídas as informações necessárias para o desenvolvimento deste trabalho.

O trabalho inicial de construção de um banco de dados facilmente utilizável e de definição da localização dos instrumentos, foi realizado por parte da equipe do projeto AIEVC – “Análise de Incertezas e Estimação de Valores de Controle para o Sistema de Monitoração Geotécnico-Estrutural na Barragem de Itaipu”, da UFPR. Também fizeram parte deste projeto, os trabalhos de Buzzi (2007) e de Sanchez (2009).

Buzzi (2007) propôs uma metodologia, baseada em correlações estatísticas, para o entendimento das interações existentes entre as séries temporais de leituras de instrumentos de monitoração geotécnico-estrutural instalados em barragens. O autor também analisou a influência que a temperatura ambiente exerce sobre as leituras destes instrumentos, visando-se estimar o atraso no tempo de resposta dos instrumentos à ocorrência de mudanças nos parâmetros ambientais, tais como picos de temperatura. A metodologia proposta possibilita identificar anomalias em leituras de instrumentos e realizar a previsão ou recuperação de leituras.

Sanchez (2009) desenvolveu e aplicou metodologias, baseadas em redes neurais artificiais e geoestatística, para a análise de leituras de instrumentos de auscultação de grandes obras de engenharia civil. O autor usou dados de piezometria da barragem de Itaipu, avaliando leituras em diferentes regiões do maciço de fundação, procurando identificar áreas mais críticas daquele sítio. Técnicas de interpolação baseadas em RNA's e geoestatística foram utilizadas para o mapeamento espacial e espaço-temporal dos níveis piezométricos de Itaipu. Segundo o autor, os modelos de predição de leituras e o mapeamento de níveis piezométricos consistem numa ferramenta útil para diagnóstico de mudanças de comportamento da barragem e podem ser incorporados aos sistemas de apoio à tomada de decisão, relacionados à segurança de barragens.

Para a maioria dos instrumentos tem-se uma leitura mensal, porém, alguns deles, apresentam mais de uma leitura por mês sendo que, nestes casos, trabalhou-se com a média mensal. Por outro lado, alguns instrumentos apresentaram leituras faltantes e, nestas situações, foram realizadas interpolações por séries temporais usando o *software Statgraphics Plus* 5.1 (2001) garantindo, desta forma, que todos os instrumentos tivessem exatamente 120 leituras. Para um melhor detalhamento a respeito de técnicas de interpolação envolvendo Séries Temporais, sugere-se o livro de Box e Jenkins, 1976.

Na interpolação por séries temporais, o modelo a ser ajustado foi escolhido segundo o critério de Akaike (AIC), observando-se, também, a raiz do erro quadrático médio (RMSE). Assim, escolhe-se o modelo cujas ordens *p* e *q* minimizam o critério descrito pela equação 3.1, onde $\delta_{d0} = \begin{cases} 1, d = 0 \\ 0, d \neq 0 \end{cases}$ (MORETTIN; TOLOI, 1985).

$$AIC(p,d,q) = N.\log\hat{\sigma}_a^2 + \frac{N}{N-d}.2.\left(p+q+12\pi+\delta_{d0}\right) + N.log2\pi + N \qquad (3.1)$$

Identificado o modelo, estimam-se os seus parâmetros pelo método de máxima verossimilhança com emprego do algoritmo de Levenberg-Marquardt. Um modelo *ARIMA* (*p*, *d*, *q*) com *d* ≥ 0 tem *p+q+1* parâmetros. Então, na metodologia Box e Jenkins, estimam-se os parâmetros do modelo, uma vez que se tenha identificado a sua estrutura ARIMA (*p*, *d*, *q*) da forma: $\phi(B)\omega_t = \theta(B)a_t$, onde

$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p$ é o polinômio característico da parte autoregressiva de ordem *p*, $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - ... - \theta_q B^q$ é o polinômio característico da parte médias móveis e $\omega_t = \nabla^d Z_t$ corresponde a série diferenciada para *d* = 0, 1, 2, ...

O procedimento de estimação do vetor de parâmetros $\underline{\eta}' = \left[\delta, \phi_1, \phi_2, ..., \theta_p, \theta_1, \theta_2, ..., \theta_q\right]$ determina o vetor $\underline{\eta}$' que minimiza a função soma de quadrados não-condicional que é dominada por $\frac{S(\eta)}{2\sigma_a^2}$ e, neste caso, corresponde a minimizar $\frac{S(\eta)}{2\sigma_a^2}$. Como se trata de minimizar quadrados não lineares deve-se aplicar um método iterativo, particularmente o Algoritmo de Marquardt.

Uma indicação de garantia do melhor modelo pode ser obtida no periodograma acumulado dos resíduos e, em alguns casos, após análise dos valores-*p* nos testes "*t*" dos parâmetros, o modelo é escolhido.

Supondo que a_t, t = 1, ..., n sejam observações do processo estocástico ruído branco ($N(0,\sigma_a^2)$), um estimador do espectro desse processo é dado pela equação (3.2), com $0 < f_i < ½$, chamado periodograma. Este estimador foi proposto com a finalidade de detectar periodicidades nos dados. O periodograma acumulado (normalizado) é dado pela equação (3.3). Para um ruído branco o gráfico de $C(f_j)$ deve estar espalhado ao redor da reta que passa pelos pontos (0,0) e (0,5;1), reta verde na figura 3.1. Para obter limites de confiança ao redor da reta, traçam-se retas a uma distância $\frac{1{,}36}{\sqrt{\frac{n-1}{2}}}$ para α = 0,05, retas vermelhas na figura 3.1 (MORETTIN; TOLOI, 1985). A figura 3.1 mostra um exemplo de periodograma acumulado.

$$I_a(f_i) = \frac{2}{n}\left[\left(\sum_{t=1}^{n} a_t \cos\frac{2\pi i}{n}t\right)^2 + \left(\sum_{t=1}^{n} a_t sen\frac{2\pi i}{n}t\right)^2\right] \tag{3.2}$$

$$C(f_j) = \frac{\sum_{i=1}^{j} I_a(f_i)}{n\hat{\sigma}_a^2} \tag{3.3}$$

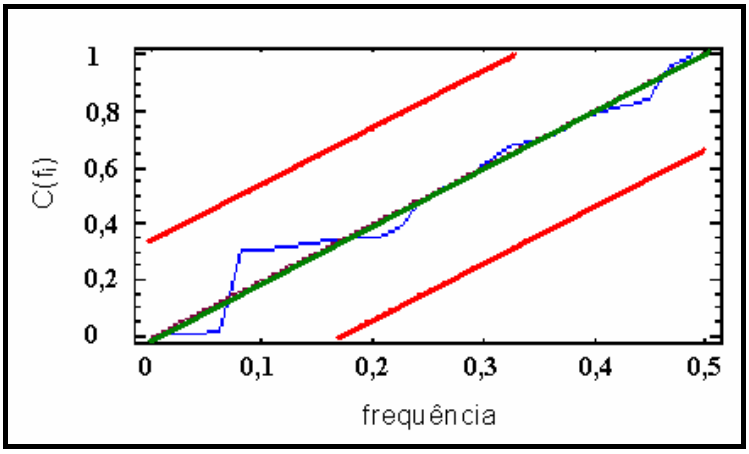

Figura 3.1 – Exemplo de Periodograma Acumulado.

Deste modo, a matriz *Q* de entrada dos dados de instrumentação geotécnica-estrutural da Itaipu é da ordem *a* x *b*, onde *a* é o número de padrões e *b* é o número de atributos. Para o caso dos dados de instrumentação geotécnicas-estrutural da Itaipu, a = 72 (número de padrões) e b = 120 (número de atributos).

Na etapa de formatação dos dados (3ª. etapa do processo *KDD*), os mesmos foram padronizados e, neste caso, a padronização foi feita por atributo. Para isso, identifica-se o valor máximo e o valor mínimo para cada atributo, subtraindo-se o valor mínimo a todos os valores e dividem-se todos os valores pelo valor máximo subtraindo-se o valor mínimo, conforme equação 3.4. Fazendo isto, todos os novos valores pertencem ao intervalo [0, 1]. Esta padronização foi feita para todas as bases de dados utilizadas.

$$x^{*}_{i,j} = \frac{x_{i,j} - x_{\min(i)}}{x_{\max(i)} - x_{\min(i)}} \tag{3.4}$$

3.4 MINERAÇÃO DE DADOS

A 4ª etapa do processo *KDD* é a Mineração de Dados (figura 2.7), sendo que neste trabalho a tarefa realizada foi a de agrupamento de padrões.

Inicialmente, apenas para a base de dados de instrumentação da Itaipu, foram aplicadas simultaneamente a Análise Fatorial (para hierarquizar as hastes de extensômetros) e, em seguida, Análise de Agrupamento (para agrupar as hastes de extensômetros semelhantes). A Análise Fatorial também foi aplicada dentro de cada

grupo formado na Análise de Agrupamento, conforme apresentado no fluxograma da figura 3.2. Os resultados são apresentados no capítulo 4.

Além da análise anterior, em Villwock *et al.* (2009), foi aplicada a Análise de Componentes Principais para selecionar as hastes de extensômetros. Considerando somente as hastes de extensômetros selecionadas foram aplicadas a Análise Fatorial (para hierarquizar as hastes) e a Análise de Agrupamento (para agrupar as hastes semelhantes). A Análise Fatorial também foi aplicada dentro de cada grupo formado pela Análise de Agrupamento. Estudos preliminares foram realizados em Villwock, Steiner e Dyminski (2007); Dyminski, Steiner e Villwock (2008) e Villwock *et al.* (2007) para fazer a hierarquização das hastes de extensômetros. Técnicas de Mineração Visual de Dados também foram utilizadas para examinar relações entre os extensômetros em Silva Neto *et al*. (2008). Destes resultados todos, apenas os mais promissores estão sendo aqui apresentados.

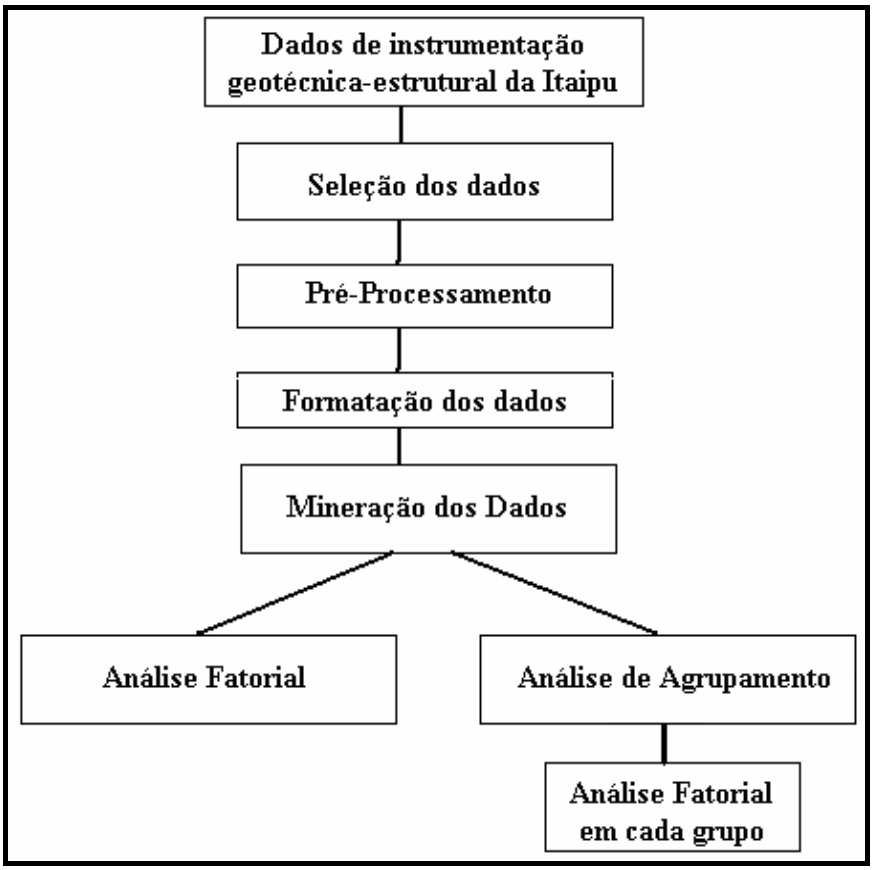

Figura 3.2 – Fluxograma mostrando as etapas do processo *KDD*, onde na etapa de Mineração de Dados foram aplicadas técnicas da Análise Multivariada dos Dados para a base de dados de Itaipu

3.4.1 Detalhamento da Aplicação da Análise Fatorial

Após a formação dos grupos, foi realizada a hierarquização dos padrões dentro de cada grupo com o auxílio da Análise Fatorial.

Na Análise Fatorial, a comunalidade h_i^2 é a porção da variância da variável que é atribuída aos fatores e representa o percentual da variação da variável que não é aleatória e sim, proveniente dos fatores. Portanto, o critério para hierarquização de padrões que foi utilizado consiste em ordenar as hastes segundo a sua comunalidade. Será avaliada a comunalidade como um fator que distingue o comportamento do instrumento, utilizando-a como um controle de qualidade prático e simples da medida do instrumento.

Para fazer a hierarquização dos atributos é utilizado o escore fatorial final, que é dado pela equação (3.5), onde m é o número de fatores, λ_i são os autovalores e f_i são os escores fatoriais.

$$escore\_fatorial\_final = \frac{\sum_{i=1}^{m} \lambda_i f_i}{\sum_{i=1}^{m} \lambda_i} \tag{3.5}$$

A Análise Fatorial foi feita com o auxílio do *software* computacional *Statgraphics Plus* 5.1 (2001).

3.4.2 Aplicação da Análise de Agrupamento através da Análise Multivariada

Nas cinco bases de dados reais e de séries temporais (IRIS, WINE, PIMA, GUN e LIGHTNING-2) utilizadas neste trabalho, a Análise de Agrupamento, através da Análise Multivariada, foi realizada com o auxílio do *software* computacional *MATLAB* R2008b (2008). Nestas cinco bases de dados é conhecido o número correto de grupos e este número foi fornecido para que o agrupamento fosse avaliado. Para isso, foram utilizados quatro métodos de ligação (já apresentados na seção 2.3.2): Ligação Simples, Ligação Completa, Ligação Média e Método Ward.

Para estas bases de dados, os métodos foram avaliados utilizando-se dois índices externos (Medida F e Índice Aleatório) e o percentual de classificação errada (conforme já apresentado na seção 2.7).

Já para a base de dados referente aos dados de instrumentação geotécnica-estrutural da Itaipu, a Análise de Agrupamento, através da Análise Multivariada, foi aplicada com o auxílio do *software* computacional *Statgraphics Plus* 5.1 (2001). A escolha deste *software* para esta base de dados se justifica pelo fato de que, neste caso, não era conhecido o número correto de grupos.

A medida de dissimilaridade utilizada foi a distância Euclidiana por ser a mais conhecida entre as medidas de dissimilaridade e por ter sido empregada em trabalhos anteriores para todos os métodos aqui utilizados. Os métodos foram avaliados utilizando-se as medidas de avaliação variância e Índice *Dunn*. Estas medidas foram utilizadas porque, para estes dados, não havia conhecimento prévio do grupo a que cada padrão pertencia.

3.5 AGRUPAMENTO DOS DADOS ATRAVÉS DAS REDES NEURAIS DE KOHONEN UNIDIMENSIONAL

O agrupamento por Redes Neurais de Kohonen Unidimensional (1D-SOM) foi aplicado às cinco bases de dados e à base de dados de instrumentação de Itaipu. Este método foi implementado no *software* computacional *MATLAB* R2008b (2008) e foi executado dez vezes.

Neste método, é necessária a definição do número de neurônios, e neste trabalho, foi definido que o número de neurônios deve ser igual ao número de grupos (*k*). O número de grupos é conhecido para as bases de dados reais e de séries temporais.

Como definido na seção 2.4.1, o primeiro passo na execução do algoritmo de Redes Neurais de Kohonen Unidimensional é a inicialização. Nesta implementação, foram definidos: a taxa de aprendizagem inicial igual a 0,5; a taxa de aprendizagem mínima igual a 0,05; o raio de vizinhança inicial igual ao valor máximo entre “1” e “¼ *k*”; os pesos sinápticos iniciais dos neurônios igual a valores aleatórios pertencentes ao intervalo [0, 1] e o número máximo de iterações *N* = 500.

O segundo passo na execução do algoritmo é a definição do critério de parada que, neste trabalho, foi definido como o número máximo de iterações. No

algoritmo implementado, foram definidas duas fases (inicial e final) na qual os ajustes de parâmetros são modificados. A fase inicial foi definida como $t_{inicial}$ = 0,2 *N*. Na fase final, o raio de vizinhança inicial é igual ao raio de vizinhança ao final da primeira fase.

O terceiro passo é o treinamento, que envolve as fases competitiva, cooperativa e adaptativa, onde cada padrão deve ser apresentado à rede. Nesta implementação, a ordem de entrada dos padrões foi definida para ser aleatória, ou seja, a cada iteração todos os padrões são apresentados à rede de forma aleatória.

Na fase competitiva, calculam-se as distâncias do padrão a todos os neurônios e verifica-se qual é o neurônio vencedor. Nesta implementação foi utilizada a distância Euclidiana.

Na fase cooperativa, localizam-se os vizinhos do neurônio vencedor e na fase adaptativa, atualizam-se os pesos sinápticos dos neurônios vizinhos ao neurônio vencedor. A atualização dos pesos sinápticos foi feita segundo a equação (2.5) (apresentada na seção 2.4.1) com função de vizinhança definida pela equação (2.6) (seção 2.4.1). Esta atualização leva em consideração a distância do vizinho até o neurônio vencedor e a taxa de aprendizagem.

No quarto passo, a taxa de aprendizagem e o raio de vizinhança devem ser atualizados, e isto foi feito segundo as equações (2.7) e (2.8) (seção 2.4.1), respectivamente. Nestas equações, $\tau_1 = \frac{N}{\log(\sigma_0)}$ e $\tau_2 = N$, onde *N* é o número máximo de iterações e σ_0 é o raio de vizinhança inicial. Estes valores foram definidos baseando-se nos valores utilizados por Haykin (2001), $\tau_1 = \frac{1000}{\log(\sigma_0)}$ e $\tau_2 = 1000$.

Para avaliação dos resultados foram utilizados dois índices externos (Medida F e Índice Aleatório) e o percentual de classificação errada.

3.6 AGRUPAMENTO DOS DADOS ATRAVÉS DO ALGORITMO BASEADO EM FORMIGAS PROPOSTO

O agrupamento baseado em Formigas proposto foi implementado no *software* computacional *MATLAB* R2008b (2008). Nesse trabalho foram utilizados recursos da grade computacional do LCPAD: Laboratório Central de Processamento de Alto Desempenho/UFPR, parcialmente financiado pela FINEP projeto CT-INFRA/UFPR/Modelagem e Computação Científica. Este algoritmo foi baseado no

algoritmo básico de Deneubourg *et al.*, apresentado na seção 2.5.7. No Anexo 3 foi desenvolvido um exemplo acadêmico do funcionamento do algoritmo.

Neste algoritmo básico de Deneubourg *et al.*, várias propostas quanto a implementação são apresentadas a seguir, com o intuito de esclarecê-lo e melhorar o seu desempenho. Alguns procedimentos permaneceram os mesmos, os quais são igualmente enfatizados. Finalizando este capítulo 3 tem-se, na seção 3.6.1, a descrição das três principais modificações propostas para o Agrupamento baseado em Formigas.

O algoritmo implementado utilizou como critério de parada o número de iterações e o algoritmo foi executado dez vezes. Sendo n é o número de padrões e m é o número de atributos, o número de iterações N_{max} foi definido como N_{max} = 500.n.m para as bases de dados reais (IRIS, WINE e PIMA) e como N_{max} = 2000.n para as bases de dados de séries temporais (GUN e LIGHTNING-2). Para definir o número máximo de iterações, vários testes foram realizados. O número máximo de iterações foi diferenciado para bases de dados de séries temporais devido ao fato destas bases possuírem números elevados de atributos. No algoritmo implementado, foram definidas duas fases (inicial e final) na qual os ajustes de parâmetros são modificados. A fase inicial foi definida como $t_{inicial}$ = 0,2.N_{max} para as bases de dados reais e como $t_{inicial}$ = 0,1.N_{max} para as bases de dados de séries temporais.

Na definição do tamanho da grade, escolheu-se o número de células igual a 10 vezes o número de padrões e foram utilizadas 10 formigas (p=10), como em Handl, Knowles e Dorigo (2006). Observou-se que a alteração destes valores não é imprescindível no processo de agrupamento e, por este motivo foram utilizados os mesmos valores. Foi utilizada vizinhança quadrada na busca dos padrões vizinhos.

Como em Handl, Knowles e Dorigo (2006), o raio de vizinhança inicial foi definido igual a “1”, com a utilização de incremento deste valor durante a fase inicial. Como não foi encontrada explicitamente em outros trabalhos da literatura uma equação para o aumento deste valor, isto foi feito segundo a equação (3.6), onde t é a iteração atual da fase inicial. Durante a fase final, este valor decresce em 0,05 a cada 100 substituições do padrão carregado por uma formiga (modificação sugerida e que será descrita na seção 3.6.1). O valor do raio de vizinhança é sempre o valor inteiro menor ou igual ao definido em qualquer uma das fases. Esta adaptação automática durante a fase final tem a finalidade de “relaxar” o tamanho da

vizinhança quando as formigas não estão conseguindo descarregar os padrões que carregam.

$$\sigma = 4^{\frac{t}{t_{inicial}}} \tag{3.6}$$

Na definição da vizinhança para o cálculo da probabilidade de descarregar um padrão em sua posição atual e para o cálculo da probabilidade de carregar um padrão, considerou-se o raio de vizinhança sempre igual a “1”.

Na busca de uma nova posição, a direção do passo é aleatória. Definida a direção, calcula-se o tamanho máximo possível do passo. Um número aleatório pertencente ao intervalo [0, 1] foi utilizado para determinar este tamanho, multiplicando-se este número pelo tamanho máximo do passo.

As probabilidades de carregar (p_{pick}) e descarregar (p_{drop}) utilizadas são as descritas pelas equações (2.9) e (2.10) (na seção 2.5.2), respectivamente, onde k_p = 0,1 e k_d = 0,3, como em Deneubourg *et al.*. Um padrão é carregado se a probabilidade p_p for maior que um valor mínimo para carregamento ($pick_{min}$). Um padrão é descarregado se a probabilidade p_d for maior que um valor mínimo para descarregamento ($drop_{min}$).

Os valores de $drop_{\min}$ e $pick_{\min}$ foram definidos como 0,13397 durante a fase inicial. Durante a fase final, estes valores foram definidos aleatoriamente, com a restrição de serem maiores que 0,13397, a cada vez que todas as formigas executassem uma iteração. O valor 0,13397 foi definido fazendo a probabilidade de carregar (p_{pick}) igual a probabilidade de descarregar (p_{drop}). A figura 3.3 mostra o gráfico das probabilidades de carregar e descarregar. A definição de um valor aleatório maior que 0,13397 durante a fase final foi feita para restringir mais a mudança de posição durante esta fase sem “engessar” o processo. A definição de um valor que aumentasse com o tempo, à longo prazo, impediria que as formigas movessem os padrões.

No cálculo da função *f*, foi utilizada a função *f** definida pela equação (2.13) (seção 2.5.2), proposta por Handl, Knowles e Dorigo (2006), substituindo-se o parâmetro escalar $\frac{1}{\sigma^2}$ por $\frac{1}{N_{occ}}$, onde N_{occ} é o número de células da grade ocupadas observadas dentro da vizinhança local. Esta escolha foi feita depois de

testes preliminares com as funções *f* originais, definidas pelas equações (2.12 e 2.13) (seção 2.5.2), propostas por Lumer e Faieta (1994, *apud* Handl, Knowles e Dorigo, 2006) e Handl, Knowles e Dorigo (2006), bem como com a substituição do parâmetro escalar $\frac{1}{\sigma^2}$ por $\frac{1}{N_{occ}}$, durante todo ou parte do processo.

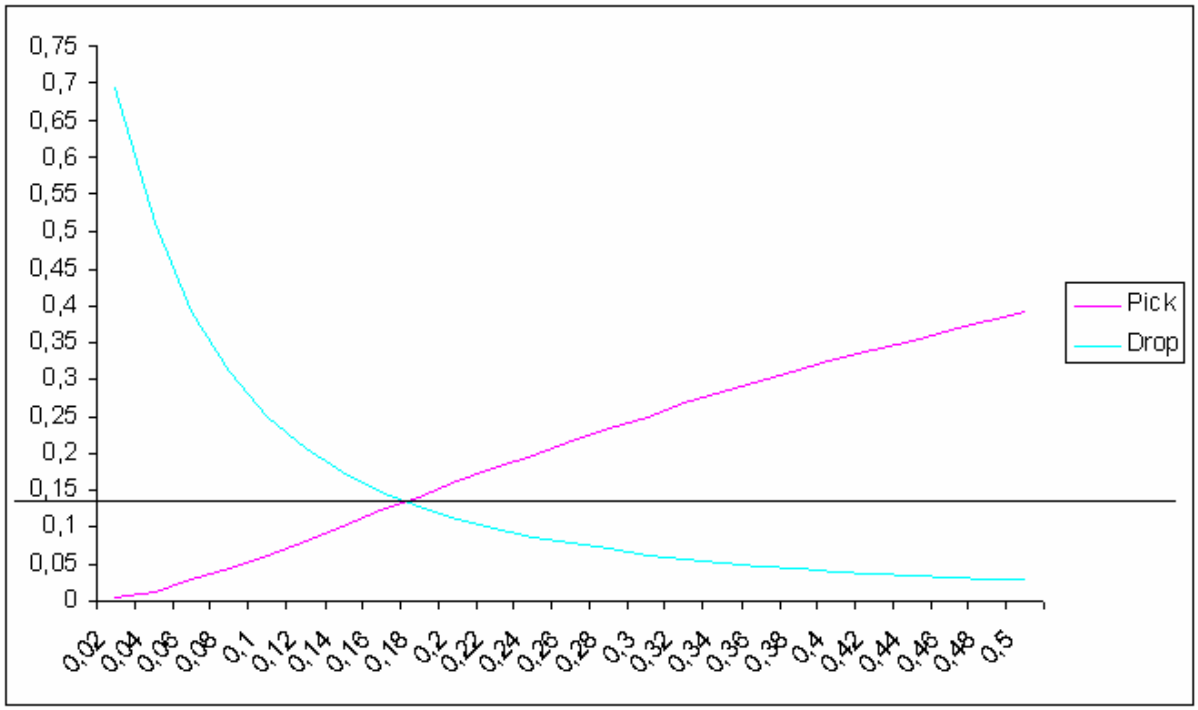

Figura 3.3 – Gráfico das probabilidades de carregar e descarregar padrões

O parâmetro *α* inicial (α_0), depois de alguns testes preliminares, foi definido como 0,8. Além destes testes, foram realizados outros testes preliminares para investigar a necessidade de atualização do valor de *α* e a forma de atualização deste valor. Baseando-se nos testes, foi definida a atualização durante a fase inicial segundo a equação (3.7). Durante a fase final este valor decresce em 0,001 a cada 100 substituições do padrão carregado por uma formiga. Este decréscimo durante a fase final foi feito para evitar que as formigas não conseguissem mais mover seus padrões.

$$\alpha = \alpha_0 + \frac{2t}{p.t_{inicial}} - 0,01 \qquad (3.7)$$

Observa-se que qualquer alteração nos valores de k_p, k_d e α influenciam diretamente o processo de agrupamento. Optou-se por manter os valores de k_p e k_d e utilizar somente uma adaptação para α. Se os valores de k_p e k_d forem alterados, a adaptação para α, bem como os valores $pick_{min}$ e $drop_{min}$ deverão ser revistos.

Quando um padrão é descarregado na grade, um novo padrão deverá ser carregado. A busca deste padrão é aleatória, porém, cada padrão livre é avaliado somente uma vez, até que todos sejam avaliados. Caso nenhum padrão apresente probabilidade p_{pick} maior que $pick_{\text{min}}$, o padrão que apresentar a maior probabilidade p_{pick} é carregado.

Quando um padrão não tem vizinhos, definiu-se a função *f* igual a zero. Isso faz com que a probabilidade p_{d} seja igual a “0”, ou seja, o padrão não deve ser descarregado naquela posição e a probabilidade p_{p} foi igual a “1”, ou seja, o dado deverá ser carregado e futuramente deixar esta posição.

A medida de dissimilaridade utilizada foi a distância Euclidiana. A matriz de distâncias foi calculada segundo a equação 3.8 e depois foi padronizada. Na equação 3.8, o *peso* se refere ao atributo e é calculado dividindo-se o desvio-padrão pela média, calculado para cada atributo da matriz dos dados já padronizada (*Q*).

$$\tilde{d}(i,j) = \sum_{a=1}^{m} \left[\left(Q(a,i) - Q(a,j) \right) . peso(a,1) \right]^2 \tag{3.8}$$

Na recuperação dos grupos foi utilizado o Método Ward e foi definido um número máximo de grupos. Em Villwock e Steiner (2008), outros métodos foram testados e o Método Ward apresentou melhores resultados.

Além dos detalhes de implementação descritos e da inclusão de melhorias já propostas anteriormente, três principais modificações foram propostas neste trabalho e são descritas a seguir.

3.6.1 Modificações Propostas para o Agrupamento baseado em Formigas

Durante o estudo do Agrupamento baseado em Formigas, foi observado que muitas das mudanças de posição dos padrões ocorrem desnecessariamente. Considera-se uma mudança desnecessária quando um padrão está entre similares na grade e, neste caso, não há necessidade da mudança deste padrão para outra

posição. Com o objetivo de evitar estas mudanças desnecessárias, uma comparação da probabilidade de descarregar um padrão na posição escolhida aleatoriamente com a probabilidade de descarregar este padrão em sua posição atual foi introduzida. A decisão de descarregar um padrão na posição escolhida aleatoriamente só ocorre, se esta probabilidade for maior que a probabilidade de descarregar este padrão em sua posição atual.

Também foi observada a ocorrência de fusão de grupos próximos na grade. Quando a decisão de descarregar um padrão for positiva e a célula em que o padrão deveria ser descarregado está ocupada, busca-se aleatoriamente uma posição vizinha a esta, que esteja livre. Porém, esta nova posição pode estar próxima também a outro grupo de padrões na grade. Este pode ser um motivo para a fusão de grupos próximos. Como uma alternativa para evitar a fusão de grupos próximos na grade, foi proposta neste trabalho uma avaliação da probabilidade para a nova posição. O padrão só é descarregado na célula vizinha se a probabilidade de descarregar o padrão nesta posição for maior que a probabilidade de descarregar este padrão em sua posição atual. Todas as posições vizinhas livres são avaliadas. Se em nenhuma posição vizinha livre a probabilidade de descarregar o padrão for maior que a probabilidade de descarregar este padrão em sua posição atual, o padrão não é descarregado e o processo se reinicia escolhendo-se outra formiga.

Outra questão observada no Agrupamento baseado em Formigas é que uma formiga pode carregar um padrão que está entre similares na grade. Uma formiga só carrega um padrão quando este não está entre similares na grade, porém, desde que a formiga carrega um padrão até ela ser sorteada para tentar descarregar o padrão, mudanças ocorrem na vizinhança deste, podendo deixá-lo então entre similares. Sendo assim, esta formiga fica inativa, pois a operação de descarregar o padrão não é executada. Neste caso, foi proposta a substituição do padrão carregado por uma formiga, caso este padrão não seja descarregado em 100 iterações consecutivas. O novo padrão foi escolhido por sorteio, mas ele só foi carregado pela formiga se a probabilidade de carregar este padrão for maior que 0,13397. O valor 0,13397 foi definido fazendo a probabilidade de carregar (p_{pick}) igual a probabilidade de descarregar (p_{drop}). Caso não exista nenhum padrão com probabilidade de carregar maior que 0,13397, o último padrão sorteado é carregado pela formiga. Este também poderia ser um critério de parada.

4 RESULTADOS E DISCUSSÃO

Inicialmente são apresentados os resultados para as cinco bases de dados (três bases de dados reais: IRIS, WINE e PIMA e duas bases de dados de séries temporais: GUN e LIGHTNING-2), através de cada uma das técnicas abordadas: Estatística Multivariada; Redes Neurais de Kohonen e Agrupamento baseado em Formigas. Na seqüência são, então, apresentados os resultados para os dados de instrumentação geotécnica-estrutural da Usina Hidrelétrica de Itaipu.

4.1 RESULTADOS DA ANÁLISE ESTATÍSTICA MULTIVARIADA

Aplicou-se os métodos de agrupamento Ligação Simples, Ligação Média, Ligação Completa e Método Ward, apresentados na seção 2.3.2, para as cinco bases de dados, nas quais as respostas (grupo a que cada padrão pertence) eram conhecidas. Os resultados da aplicação dos métodos são apresentados nos quadros 4.1 a 4.5, a seguir.

IRIS	Ward	Ligação Simples	Ligação Completa	Ligação Média
R (quanto maior, melhor)	**0,957**	0,777	0,950	**0,957**
F (quanto maior, melhor)	**0,967**	0,775	0,960	**0,967**
Classificação errada (%)	**3,333**	32,000	4,000	**3,333**

Quadro 4.1 – Resultados da aplicação dos métodos de agrupamento, através da Análise Multivariada, para a base de dados IRIS.

WINE	Ward	Ligação Simples	Ligação Completa	Ligação Média
R (quanto maior, melhor)	0,819	0,343	**0,870**	0,716
F (quanto maior, melhor)	0,845	0,505	**0,898**	0,736
Classificação errada (%)	15,169	61,236	**10,112**	36,517

Quadro 4.2 – Resultados da aplicação dos métodos de agrupamento, através da Análise Multivariada, para a base de dados WINE.

PIMA	Ward	Ligação Simples	Ligação Completa	Ligação Média
R (quanto maior, melhor)	0,531	**0,546**	0,538	0,545
F (quanto maior, melhor)	0,624	**0,694**	0,646	0,693
Classificação errada (%)	37,370	**34,766**	36,198	34,896

Quadro 4.3 – Resultados da aplicação dos métodos de agrupamento, através da Análise Multivariada, para a base de dados PIMA.

GUN	Ward	Ligação Simples	Ligação Completa	Ligação Média
R (quanto maior, melhor)	0,497	0,497	0,501	**0,505**
F (quanto maior, melhor)	0,500	0,500	**0,658**	0,653
Classificação errada (%)	50,000	50,000	46,000	**44,000**

Quadro 4.4 – Resultados da aplicação dos métodos de agrupamento, através da Análise Multivariada, para a base de dados GUN.

LIGHTNING-2	Ward	Ligação Simples	Ligação Completa	Ligação Média
R (quanto maior, melhor)	**0,633**	0,525	0,616	0,525
F (quanto maior, melhor)	**0,739**	0,677	0,724	0,677
Classificação errada (%)	**23,967**	38,017	25,620	38,017

Quadro 4.5 – Resultados da aplicação dos métodos de agrupamento, através da Análise Multivariada, para a base de dados LIGHTNING-2.

A Ligação Simples foi o melhor método para a base de dados PIMA e a Ligação Completa, para a base de dados WINE. A Ligação Média foi o melhor método para a base de dados GUN e o Método Ward, para a base de dados LIGHTINING-2 e houve empate entre Ligação Média e Método Ward para a base de dados IRIS. Desta forma, considerou-se que o melhor método foi o Método Ward, pois este foi o melhor método em duas bases de dados, assim como a Ligação Média e, ainda, nas bases de dados onde este método foi melhor, ele foi melhor em todas as medidas de avaliação (quadros 4.1 e 4.5), apresentando um percentual de erros médio para as cinco bases menor do que para a Ligação Média.

Por este motivo, na Análise de Agrupamento aplicada aos dados de instrumentação geotécnica-estrutural da Itaipu, o método utilizado foi o Método Ward. Nesta aplicação, relembrando, os padrões são as hastes dos extensômetros e suas leituras ao longo dos meses são comparadas para a determinação dos grupos. O dendrograma da figura 4.1 mostra a formação dos grupos para estes dados.

Observando o primeiro corte, resultam dois grupos. O primeiro grupo, aqui denominado “grupo 1”, é formado por hastes de extensômetros consideradas extremamente importantes para o monitoramento da barragem. São hastes de extensômetros instalados no eixo do bloco, à montante da barragem e inclinados 60º à montante.

Observando o segundo corte, tem-se a formação de dois grupos adicionais. O primeiro, denominado “grupo 2”, possui a maioria das hastes de extensômetros instaladas nos derrames basálticos B, C e D e (A e B são camadas de rochas mais

profundas; C e D são camadas de rochas superficiais) nos contatos litológicos B/C e C/D. O segundo grupo, denominado “grupo 3”, possui a maioria das hastes de extensômetros instaladas nas juntas (entre camadas rochosas) A e B e no contato litológico A/B.

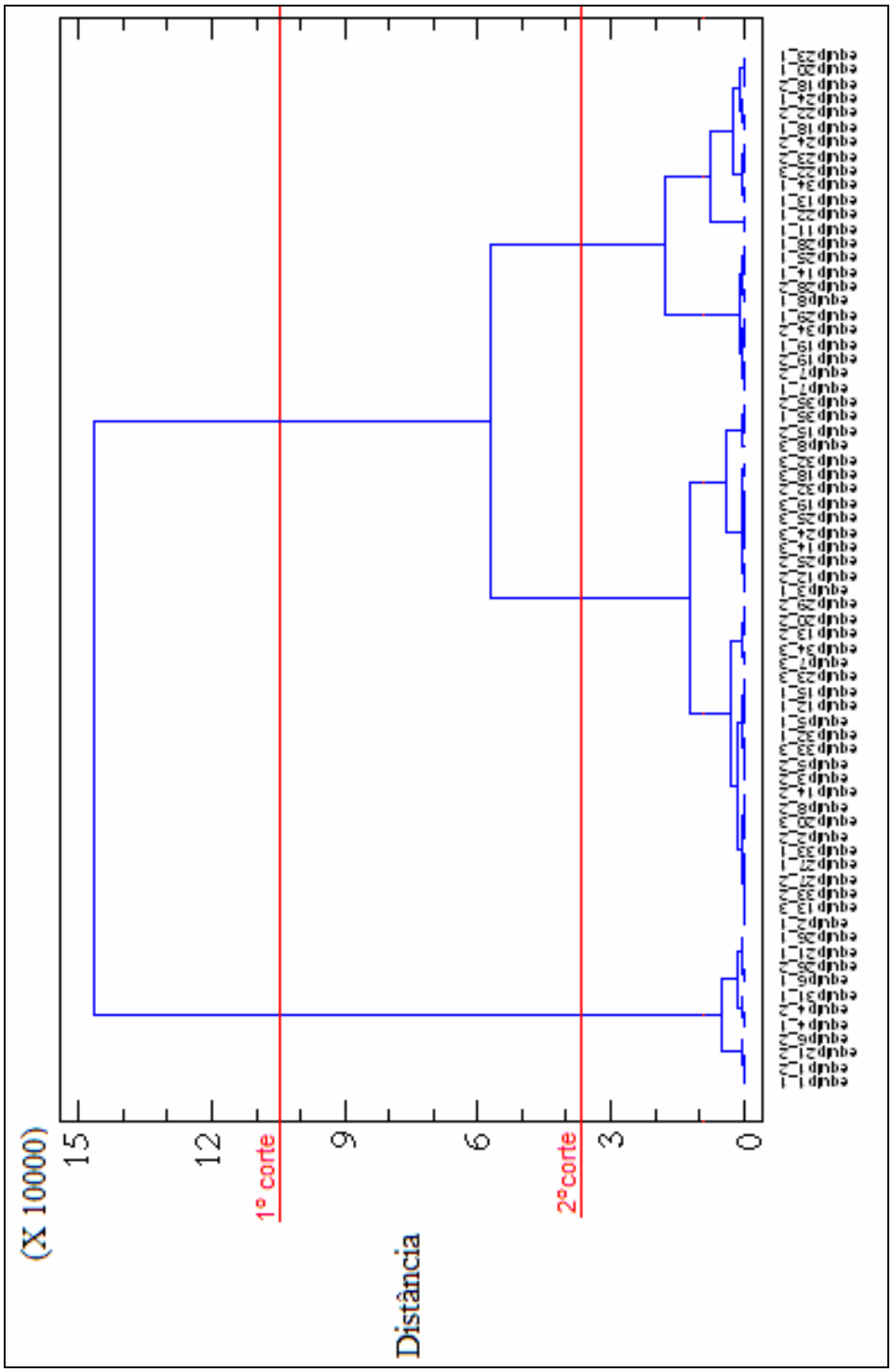

Figura 4.1 – Dendrograma mostrando a formação dos grupos em cortes diferentes (Método Ward).

Esta foi a quantidade de grupos considerada (3 grupos), já que obteve-se justificativas técnicas para sua formação. Numa subdivisão maior, não foi observada tal justificativa.

O quadro 4.6 mostra as hastes de extensômetros e a sua classificação em cada um dos três grupos, a inclinação, o afastamento dos mesmos em relação ao eixo da barragem e a feição onde a haste está instalada, conforme dendograma da figura 4.1. A figura 2 do Anexo 2 mostra os derrames de “A” a “E” do perfil basáltico do maciço de fundação da Itaipu e a figura 2.5 (seção 2.1) mostra um perfil geológico típico do maciço de fundação do trecho sem túnel da Barragem Lateral Direita da Itaipu, onde podem-se observar as principais descontinuidades (contatos, brechas e juntas) daquele sitio.

Nota-se aqui que se conseguiu agrupar os instrumentos segundo as características geológicas relevantes do maciço de fundação, mesmo não tendo as mesmas sido explicitamente apresentadas aos métodos de mineração de dados. Porém, no grupo 2, foram observadas três hastes de extensômetros instaladas na junta B e, no grupo 3, foram observadas três hastes de extensômetros instaladas nos derrames basálticos B e C e no contato litológico B/C. Estas seis hastes encontram-se em negrito no quadro 4.6 a seguir.

Grupo	Haste	Inclinação	Afastamento do eixo da barragem	Feição
1	equip 1_1	60° à M	125,5 m à M	Junta B
1	equip 1_2	60° à M	105,4 m à M	Contato B_C
1	equip 4_1	60° à M	65,3 m à M	Contato C_D
1	equip 4_2	60° à M	60,4 m à M	Rocha Fraturada
1	equip 6_1	60° à M	150,8 m à M	Junta A
1	equip 6_2	60° à M	110,5 m à M	Derrame B
1	equip 21_1	60° à M	159,8 m à M	Junta A
1	equip 21_2	60° à M	135,1 m à M	Derrame B
1	equip 26_1	60° à M	139,2 m à M	Junta B
1	equip 26_2	60° à M	115,6 m à M	Contato B_C
1	equip 31_1	60° à M	64,7 m à M	Contato C_D
2	equip 2_1	0	32,0m à M	Contato C_D
2	equip 2_2	0	32,0 m à M	Rocha Fraturada
2	equip 3_1	0	32,0 m à M	Derrame C
2	equip 3_2	0	32,0 m à M	Derrame D
2	equip 5_1	0	13,0 m à J	Contato C_D
2	equip 5_2	0	13,0 m à J	Derrame D
2	equip 7_3	0	13,0 m à J	Derrame B
2	equip 8_2	0	84,0 m à M	Rocha Fraturada
2	equip 8_3	0	84,0 m à M	Derrame B
2	equip 12_1	60° à J	47,2 m à J	Rocha Fraturada
2	equip 12_2	60° à J	42,5 m à J	Basalto Denso

(segue)

(continuação)

Grupo	Haste	Inclinação	Afastamento do eixo da barragem	Feição
2	equip 13_2	0	44,0 m à J	Rocha Fraturada
2	equip 13_3	0	44,0 m à J	Derrame B
2	equip 14_2	0	54,0 m à J	Rocha Fraturada
2	equip 14_3	0	54,0 m à J	Derrame B
2	equip 15_1	0	80,0 m à M	Rocha Fraturada
2	equip 15_2	0	80,0 m à M	Derrame B
2	equip 18_3	0	33,0 m à J	Derrame B
2	equip 19_3	0	55,0 m à J	Derrame B
2	equip 20 _2	0	82,0 m à M	Rocha Fraturada
2	equip 20_3	0	82,0 m à M	Derrame B
2	equip 23_3	0	36,0 m à J	Rocha Fraturada
2	equip 24_3	0	62,0 m à J	Derrame B
2	equip 25_2	0	75,0 m à M	Derrame B
2	equip 25_3	0	75,0 m à M	Derrame B
2	equip 27_1	30° à M	16,6 m à J	**Junta B**
2	equip 27_2	30° à M	22,6 m à J	Contato B_C
2	equip 29_2	30° à J	55,7 m à J	Contato B_C
2	equip 32_1	30° à M	36,5 m à M	**Junta B**
2	equip 32_2	30° à M	14,6 m à M	Derrame C
2	equip 32_3	30° à M	7,5 m à M	Contato C_D
2	equip 33_1	0	0,0	**Junta B**
2	equip 33_2	0	0,0	Derrame C
2	equip 33_3	0	0,0	Contato C_D
2	equip 34_3	30° à J	7,5 m à J	Contato C_D
2	equip 35_1	90° à M	0,0	Concreto
2	equip 35_2	90° à M	0,0	Concreto
3	equip 7_1	0	13,0 m à J	Junta A
3	equip 7_2	0	13,0 m à J	Contato A_B
3	equip 8_1	0	84,0 m à M	Contato A_B
3	equip 11_1	0	81,0 m à M	Junta A
3	equip 13_1	0	44,0 m à J	Contato A_B
3	equip 14_1	0	54,0 m à J	Contato A_B
3	equip 18_1	0	33,0 m à J	Junta A
3	equip 18_2	0	33,0 m à J	Rocha Fraturada
3	equip 19_1	0	55,0 m à J	Junta A
3	equip 19_2	0	55,0 m à J	Rocha Fraturada
3	equip 20_1	0	82,0 m à M	Rocha Fraturada
3	equip 22_1	0	68,0 m à M	Junta A
3	equip 22_2	0	68,0 m à M	Rocha Fraturada
3	equip 22_3	0	68,0 m à M	**Derrame B**
3	equip 23_1	0	36,0 m à J	Junta A
3	equip 23_2	0	36,0 m à J	Rocha Fraturada
3	equip 24_1	0	62,0 m à J	Junta A
3	equip 24_2	0	62,0 m à J	Rocha Fraturada
3	equip 25_1	0	75,0 m à M	Junta A
3	equip 28_1	0	40,0 m à J	Junta B
3	equip 28_2	0	40,0 m à J	**Contato B_C**
3	equip 29_1	30° à J	63,5 m à J	Junta B
3	equip 34_1	30° à J	36,6 m à J	Junta B
3	equip 34_2	30° à J	21,0 m à J	**Derrame C**

Quadro 4.6 – Classificação das hastes dos extensômetros em cada um dos três grupos, conforme dendograma da figura 4.1.

As figuras 4.2, 4.3 e 4.4, a seguir, mostram os gráficos das hastes de extensômetros nos grupos 1, 2 e 3, respectivamente. Pode-se observar a semelhança de comportamento entre estas hastes, como era de se esperar.

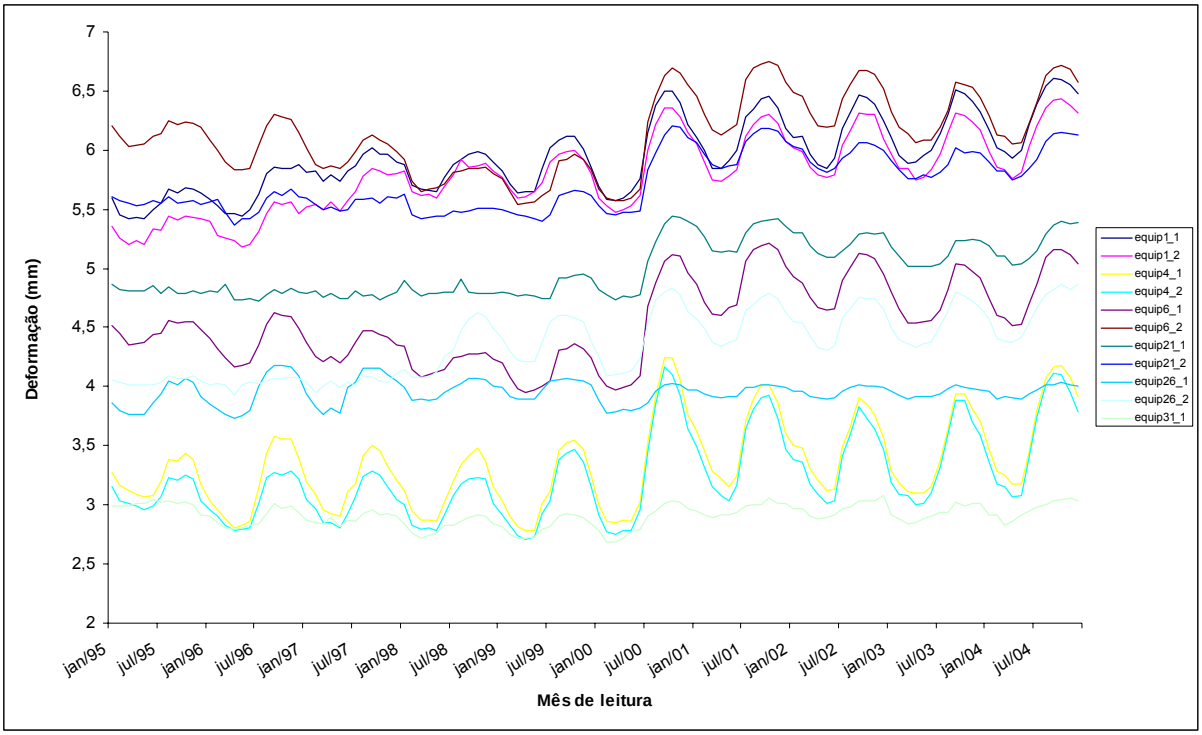

Figura 4.2 – Gráfico das hastes de extensômetros do grupo 1.

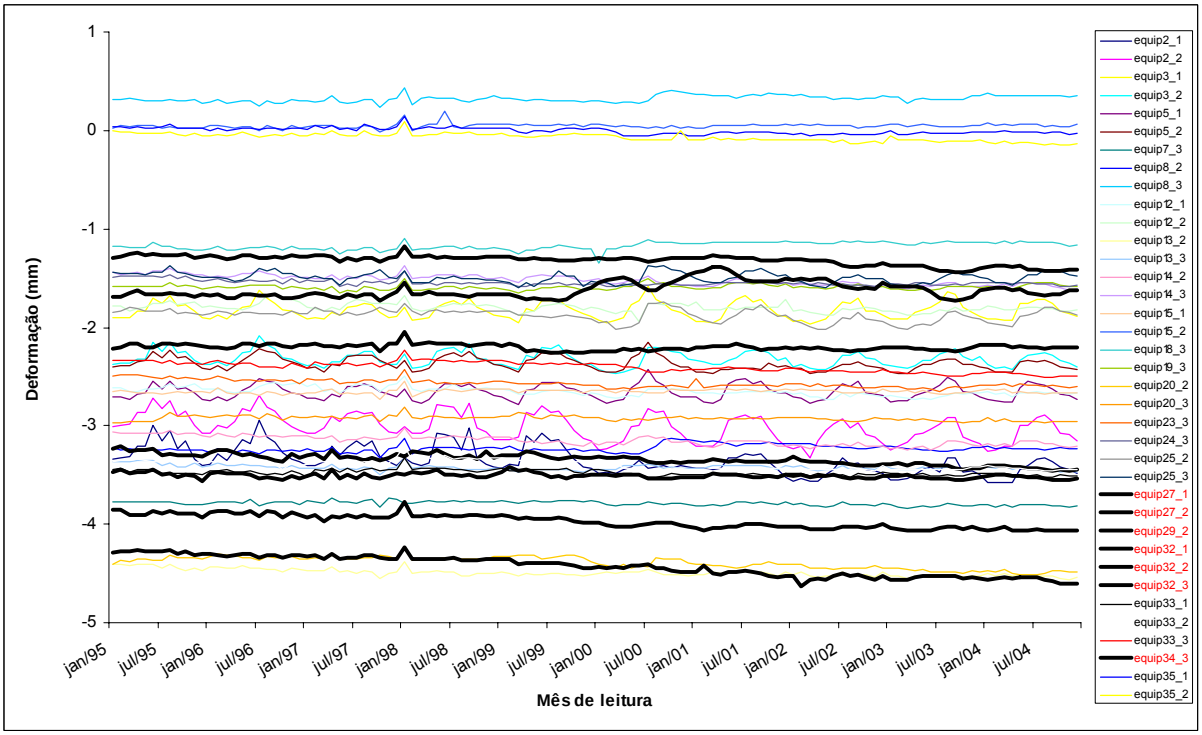

Figura 4.3 – Gráfico das hastes de extensômetros do grupo 2.

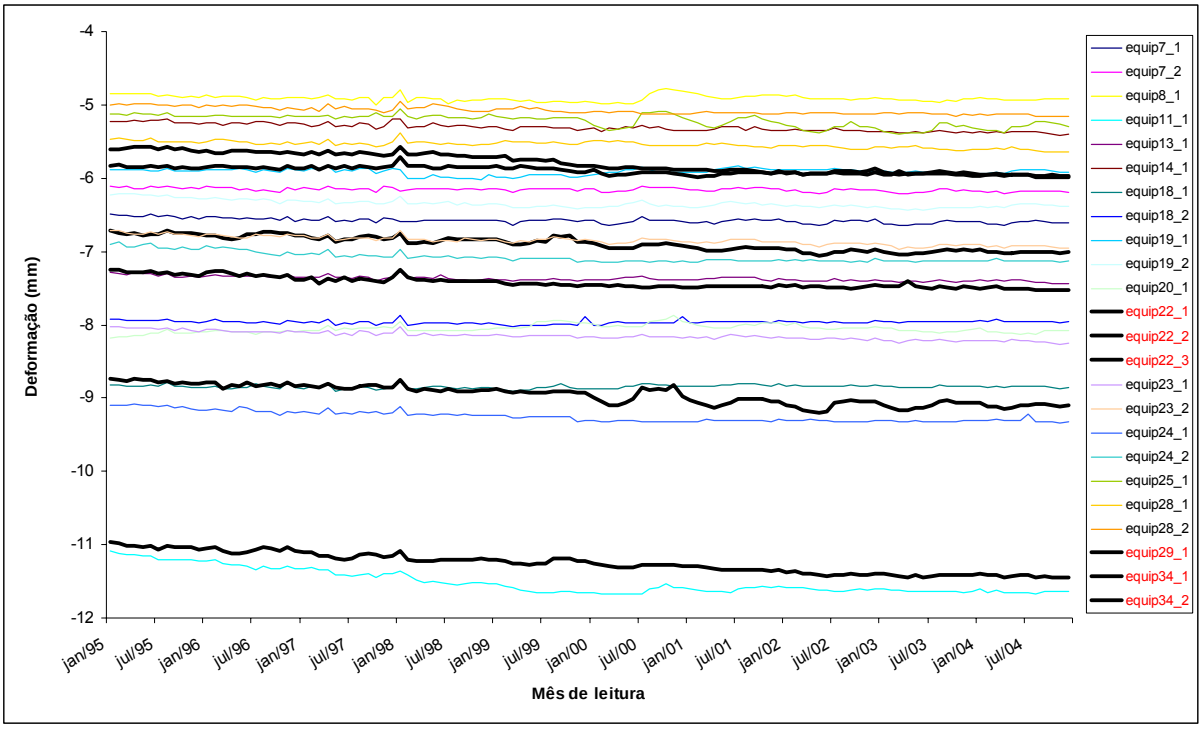

Figura 4.4 – Gráfico das hastes de extensômetros do grupo 3.

No grupo 1, todas as hastes foram automatizadas pela Itaipu. Nos grupos 2 e 3, as hastes automatizadas estão identificadas nos gráficos com a linha mais forte. Observa-se uma boa distribuição das hastes automatizadas.

A figura 4.5 mostra o gráfico de todas as hastes de extensômetros no período estudado. As linhas foram coloridas conforme o grupo à que as hastes pertencem (preto, azul e amarelo para grupos 1, 2 e 3, respectivamente). Pode-se observar a distinção entre os grupos. Deve-se salientar que esta distinção de grupos não é facilmente observada sem o conhecimento prévio destes três grupos. Analisando-se um conjunto maior de dados, esta tarefa ficaria inviável, daí a importância deste tipo de análise.

O grupo 1, composto por hastes de extensômetros instaladas à montante da barragem, mostra claramente os efeitos do verão e do inverno ilustrados na figura 2.2 (seção 2.1). Os grupos 2 e 3 são separados em função das medidas absolutas. Esta separação pode ser justificada pelo fato de estarem em feições distintas, mais superficiais no caso do grupo 2 e mais profundas no caso do grupo 3. Como as leituras das hastes mais profundas somam-se as leituras das hastes mais superficiais, estas medidas são maiores.

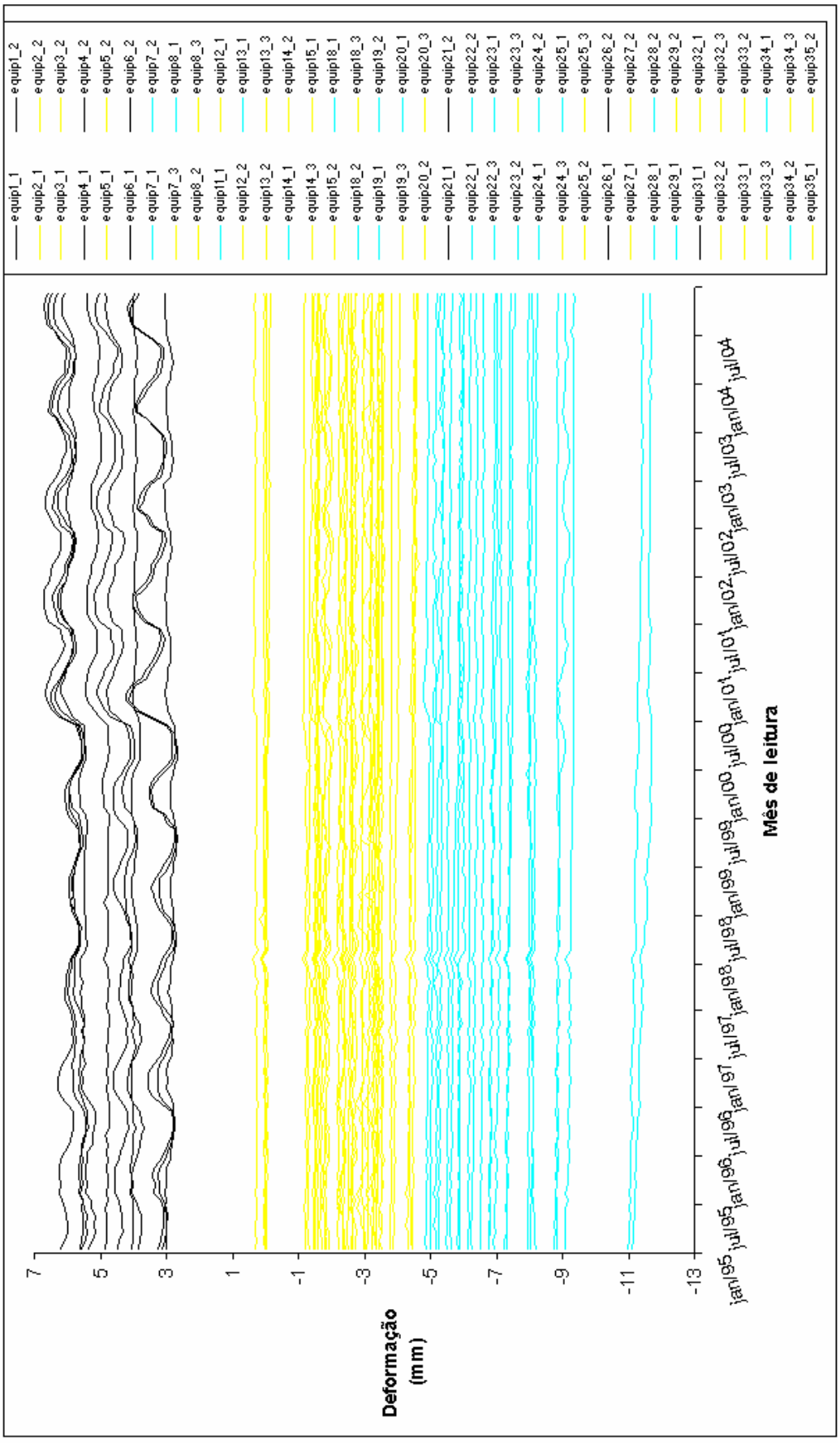

Figura 4.5 – Gráfico de todas as hastes de extensômetros no período estudado.

Foi realizada a hierarquização das 72 hastes com o auxílio da Análise Fatorial conforme figura 3.2 (seção 3.4). Nesta aplicação, foi utilizado o critério de Kaiser e verificou-se que oito fatores explicam 90,33% da variabilidade total. Como cada fator é uma combinação linear das variáveis, que são as hastes de extensômetros, cada haste tem uma contribuição para aquele fator e algumas hastes são mais importantes para um determinador fator do que para os outros.

O quadro 4.7, a seguir, mostra os pesos para cada uma das hastes em cada um dos oito fatores. Os pesos identificados em negrito são os maiores para cada haste, eles indicam quais hastes dominam cada fator.

hastes	Pesos							
	Fator 1	**Fator 2**	**Fator 3**	**Fator 4**	**Fator 5**	**Fator 6**	**Fator 7**	**Fator 8**
equip1_1	-0,541	**0,785**	0,162	-0,115	0,093	0,051	-0,096	-0,053
equip1_2	-0,595	**0,712**	0,120	-0,185	0,193	0,128	-0,043	-0,068
equip2_1	0,535	-0,206	**0,693**	-0,083	-0,052	-0,038	0,022	0,039
equip2_2	0,386	-0,076	**0,850**	0,029	-0,129	-0,093	0,110	0,059
equip3_1	-0,004	0,139	**0,964**	0,114	-0,035	-0,084	-0,025	-0,046
equip3_2	0,239	0,308	**0,848**	0,030	0,011	0,033	-0,053	0,041
equip4_1	-0,235	**0,924**	0,170	-0,007	-0,089	0,036	-0,132	-0,051
equip4_2	-0,292	**0,913**	0,182	0,005	-0,067	0,024	-0,050	-0,078
equip5_1	0,012	0,145	**0,947**	0,118	-0,033	0,011	-0,080	-0,095
equip5_2	0,298	-0,012	**0,863**	-0,050	-0,020	-0,054	-0,164	-0,075
equip6_1	-0,353	**0,863**	-0,004	0,282	-0,120	0,002	-0,019	0,033
equip6_2	-0,239	**0,880**	-0,004	0,308	-0,187	-0,015	-0,042	0,023
equip7_1	**0,811**	0,165	0,269	0,206	-0,056	-0,104	0,143	-0,313
equip7_2	**0,696**	0,230	0,224	0,085	0,029	0,113	0,365	-0,361
equip7_3	**0,651**	-0,111	0,226	-0,239	0,319	0,087	0,129	-0,349
equip8_1	0,456	**0,512**	-0,215	0,361	0,035	0,297	0,301	0,213
equip8_2	-0,130	0,526	-0,204	0,236	0,359	**0,560**	0,177	0,200
equip8_3	-0,230	**0,607**	-0,157	0,212	0,471	0,311	0,226	0,234
equip11_1	**0,887**	-0,070	-0,076	0,213	-0,227	-0,248	0,030	0,120
equip12_1	**0,803**	0,199	0,264	-0,061	0,086	0,060	-0,225	0,239
equip12_2	**0,619**	0,328	0,330	0,010	0,005	0,152	-0,421	0,087
equip13_1	**0,793**	-0,218	0,151	0,459	0,072	0,040	0,127	0,002
equip13_2	**0,811**	-0,206	0,097	0,450	0,032	0,003	0,165	0,007
equip13_3	0,530	-0,055	0,107	**0,690**	0,176	-0,020	0,302	0,045
equip14_1	**0,876**	-0,250	0,248	0,205	-0,044	0,035	0,021	0,023
equip14_2	**0,854**	-0,128	0,284	0,119	0,006	-0,094	0,111	0,161
equip14_3	**0,835**	-0,305	0,395	0,053	0,087	-0,016	-0,037	-0,004
equip15_1	0,024	-0,025	-0,193	0,236	**0,740**	0,419	-0,008	0,016
equip15_2	-0,080	-0,067	-0,013	0,127	**0,838**	0,054	0,026	0,082
equip18_1	0,082	0,423	0,295	**0,711**	0,216	0,138	-0,042	0,016
equip18_2	0,155	0,264	-0,159	**0,721**	0,288	-0,176	-0,041	0,026
equip18_3	-0,478	0,462	0,174	**0,541**	0,234	0,064	0,068	0,054
equip19_1	0,066	0,231	0,059	**0,741**	-0,310	0,050	0,054	-0,017
equip19_2	**0,831**	-0,047	0,181	0,389	-0,090	-0,176	0,130	0,096
equip19_3	0,066	0,411	0,509	**0,586**	0,189	0,070	0,059	0,013

(segue)

(continuação)

hastes	Pesos							
	Fator 1	Fator 2	Fator 3	Fator 4	Fator 5	Fator 6	Fator 7	Fator 8
equip20_1	-0,171	0,190	0,020	-0,197	0,195	**0,820**	-0,098	-0,284
equip20_2	**0,818**	-0,199	0,083	-0,293	0,057	0,340	-0,060	-0,029
equip20_3	**0,512**	-0,189	0,033	-0,134	0,287	0,499	-0,233	0,146
equip21_1	-0,550	**0,749**	-0,104	0,204	0,038	0,180	0,132	-0,003
equip21_2	-0,507	**0,798**	-0,076	0,241	-0,030	0,145	0,047	-0,003
equip22_1	**0,947**	-0,217	0,070	-0,004	-0,143	-0,002	-0,075	-0,021
equip22_2	**0,940**	-0,045	0,204	-0,152	-0,080	-0,015	0,039	-0,044
equip22_3	**0,940**	-0,137	0,170	-0,100	-0,069	-0,004	-0,152	-0,047
equip23_1	**0,940**	-0,195	0,120	0,190	-0,040	0,023	0,040	-0,033
equip23_2	**0,939**	-0,061	0,168	0,155	-0,056	0,113	-0,003	-0,120
equip23_3	**0,907**	-0,081	-0,045	0,159	0,109	-0,136	0,007	0,130
equip24_1	**0,896**	-0,229	0,065	0,132	-0,024	-0,245	-0,003	0,110
equip24_2	**0,861**	-0,179	-0,014	0,279	-0,083	-0,294	0,068	0,018
equip24_3	**0,734**	-0,015	-0,011	0,509	0,149	-0,199	-0,049	0,112
equip25_1	**0,848**	0,181	0,323	-0,206	-0,018	0,153	0,176	0,008
equip25_2	**0,612**	0,451	0,436	-0,195	0,015	-0,019	0,287	0,113
equip25_3	0,398	**0,640**	0,522	0,174	-0,136	-0,042	-0,030	0,018
equip26_1	0,031	0,504	0,272	-0,299	0,020	0,103	**-0,663**	0,035
equip26_2	-0,607	**0,675**	0,128	-0,112	0,217	0,110	0,042	-0,147
equip27_1	0,404	-0,485	-0,143	0,029	0,320	0,053	0,447	-0,080
equip27_2	**0,767**	-0,504	-0,048	-0,032	0,045	0,124	0,195	-0,150
equip28_1	**0,720**	-0,550	0,007	0,137	0,117	0,265	0,106	-0,093
equip28_2	**0,819**	-0,442	0,133	0,092	0,056	-0,063	0,107	0,126
equip29_1	**0,913**	-0,336	0,132	-0,100	-0,058	-0,057	-0,002	0,092
equip29_2	**0,897**	-0,346	0,173	-0,078	-0,068	-0,002	0,008	0,044
equip31_1	0,051	**0,828**	0,088	0,329	-0,145	-0,169	-0,131	-0,070
equip32_1	0,481	-0,045	-0,057	0,104	0,331	-0,075	-0,044	**0,718**
equip32_2	-0,401	0,073	-0,315	0,438	0,021	**0,630**	0,097	0,123
equip32_3	**0,753**	-0,341	-0,058	0,075	0,057	0,474	0,048	0,113
equip33_1	**0,595**	-0,361	-0,025	-0,201	0,543	0,033	-0,030	-0,020
equip33_2	**0,732**	-0,363	0,019	-0,255	0,447	-0,042	-0,007	0,123
equip33_3	**0,785**	-0,452	-0,007	-0,148	0,254	0,168	0,011	-0,008
equip34_1	**0,898**	-0,254	0,000	0,163	-0,042	-0,190	-0,028	0,074
equip34_2	**0,828**	-0,215	0,168	-0,153	0,238	-0,112	-0,032	0,015
equip34_3	**0,927**	-0,236	0,159	-0,119	0,083	-0,045	-0,033	0,038
equip35_1	**0,801**	-0,197	0,074	-0,103	0,331	-0,166	-0,203	0,148
equip35_2	**0,801**	-0,344	0,018	-0,100	0,219	0,173	-0,113	0,010

Quadro 4.7 – Pesos das hastes de extensômetros para cada fator.

No quadro 4.8 apresentam-se as hastes importantes para cada um dos oito fatores, ou seja, as hastes que dominam cada fator. Observa-se no quadro 4.8 que o fator 2 é dominado pelas hastes equip1_1, equip1_2, equip4_1, equip4_2, equip6_1, equip6_2, equip8_1, equip8_3, equip21_1, equip21_2, equip25_3, equip26_2 e equip31_1. Este fator contempla 10 das 11 hastes que compõem o grupo 1, confirmando que há um fenônemo externo que os influenciam. Como já mencionado, estas hastes refletem os efeitos do verão e do inverno. Da mesma forma, pode-se

observar que cada fator é dominado por um conjunto de hastes e para cada conjunto de hastes, ou para cada fator, existe um fenômeno externo que os explicam, embora estes não sejam facilmente interpretados.

fator1	fator2	fator3	fator4	Fator5	fator6	fator7	fator8
equip7_1	equip1_1	equip2_1	equip13_3	equip15_1	equip8_2	equip26_1	equip32_1
equip7_2	equip1_2	equip2_2	equip18_1	equip15_2	equip20_1		
equip7_3	equip4_1	equip3_1	equip18_2		equip32_2		
equip11_1	equip4_2	equip3_2	equip18_3				
equip12_1	equip6_1	equip5_1	equip19_1				
equip12_2	equip6_2	equip5_2	equip19_3				
equip13_1	equip8_1						
equip13_2	equip8_3						
equip14_1	equip21_1						
equip14_2	equip21_2						
equip14_3	equip25_3						
equip19_2	equip26_2						
equip20_2	equip31_1						
equip20_3							
equip22_1							
equip22_2							
equip22_3							
equip23_1							
equip23_2							
equip23_3							
equip24_1							
equip24_2							
equip24_3							
equip25_1							
equip25_2							
equip27_2							
equip28_1							
equip28_2							
equip29_1							
equip29_2							
equip32_3							
equip33_1							
equip33_2							
equip33_3							
equip34_1							
equip34_2							
equip34_3							
equip35_1							
equip35_2							

Quadro 4.8 – Hastes de extensômetros importantes para cada fator, conforme os pesos apresentados no quadro 4.7.

A comunalidade é a porção da variância das hastes de extensômetros que é explicada pelos fatores. A comunalidade baixa de uma haste indica que aquela

haste não é influenciada fortemente pelos fatores, já que a comunalidade é a soma da contribuição de cada haste para cada fator ao quadrado. Portanto, neste caso, a influência vem do fator aleatório. Observou-se que nenhuma haste de extensômetro apresentou comunalidade menor que 0,71, ou seja, nenhuma tem variação aleatória maior que 29%. A comunalidade igual a 0,71 significa que 71% da variância da haste de extensômetro é atribuída aos fatores e que somente 29% da variância é aleatória, ou seja, as hastes correspondentes funcionam bem. A comunalidade baixa indicaria a investigação da haste.

No quadro 4.9 apresentam-se as 25 hastes de extensômetros com as comunalidades mais altas. Em caso de intensificação de leituras, estas hastes são as recomendadas. As hastes destacadas fazem parte do sistema de aquisição automática de dados da Itaipu. Das 72 hastes analisadas 24 foram automatizadas pela equipe de engenheiros da Itaipu. Assim, o método de hierarquização proposto (sem o prévio agrupamento das hastes) identificou 14 das 24 hastes automatizadas.

Comunalidade	Haste
0,988861	**equip29_1**
0,981763	**equip21_2**
0,976523	equip23_1
0,975655	**equip22_1**
0,972231	equip3_1
0,971971	**equip1_1**
0,971798	**equip22_3**
0,970804	equip11_1
0,970397	**equip1_2**
0,968213	equip23_2
0,968083	**equip4_1**
0,967029	**equip21_1**
0,966632	**equip4_2**
0,965999	**equip29_2**
0,965522	**equip34_3**
0,964925	**equip6_1**
0,963139	**equip6_2**
0,960121	**equip22_2**
0,957609	equip14_3
0,953036	equip25_1
0,950395	equip33_2
0,949394	equip24_2
0,949108	equip24_1
0,948646	equip5_1
0,943644	equip28_1

Quadro 4.9 – As 25 hastes de extensômetros com as comunalidades mais altas.

Após a formação dos três grupos, foi realizada a hierarquização das hastes dentro de cada grupo com o auxílio da Análise Fatorial, conforme figura 3.2 (seção 3.4). Os quadros 4.10, 4.11 e 4.12 mostram as hastes de extensômetros e suas comunalidades para os grupos 1, 2 e 3, respectivamente.

A hierarquização dentro de cada grupo também pode ser utilizada para indicar hastes para intensificação de leituras. A vantagem da aplicação da hierarquização dentro do grupo é que se obtém primeiro uma separação das hastes com comportamentos semelhantes, sendo que as hastes indicadas representam bem a variabilidade do grupo. As hastes de extensômetros identificadas em negrito são as automatizadas pela Itaipu. Observou-se que as hastes de extensômetros automatizados estão, na maioria das vezes, entre as primeiras do *ranking* em cada grupo.

Como já mencionado acima, a comunalidade baixa de uma haste indica que aquela haste não é influenciada fortemente pelos fatores e, neste caso, a influência vem do fator aleatório. Na aplicação da análise fatorial dentro de cada um dos grupos, observou-se hastes de extensômetro com comunalidades entre 0,6 e 0,7, ou seja, variação aleatória entre 30 e 40%. Neste caso, indica-se a investigação das hastes.

Além disso, com o intuito de identificar as 24 hastes mais relevantes, optou-se por identificar as oito hastes melhores ranqueadas em cada grupo. Neste caso, obter-se-ia 15 das 24 hastes automatizadas. Este número de hastes coincidentes com as automatizadas pela Itaipu aumentaria com o auxílio de um especialista para uma melhor interpretação dos resultados. Este especialista detectaria que o grupo 1, por exemplo, é formado por hastes extremamente importantes no monitoramento da barragem e que todas as hastes deste grupo deveriam ser automatizadas.

Comunalidade	**Haste**
0,961839	**equip21_1**
0,956979	**equip21_2**
0,953791	**equip4_1**
0,94278	**equip4_2**
0,911982	**equip6_1**
0,885099	**equip1_1**
0,881852	**equip26_1**
0,854339	**equip6_2**
0,809062	**equip1_2**
0,798566	**equip26_2**
0,677401	**equip31_1**

Quadro 4.10 – Hastes de extensômetros e suas comunalidades – Grupo 1.

Comunalidade	Haste
0,958451	equip3_1
0,957903	**equip29_2**
0,956985	**equip34_3**
0,949573	equip14_3
0,942628	equip33_3
0,938697	equip33_2
0,930857	**equip32_3**
0,92853	equip5_1
0,928364	**equip27_2**
0,925976	equip13_2
0,91628	equip20_2
0,905482	equip12_1
0,903051	equip13_3
0,899792	equip35_1
0,896858	equip2_2
0,895818	equip3_2
0,895145	equip8_2
0,894009	equip14_2
0,890046	equip23_3
0,888025	equip25_3
0,859488	equip5_2
0,854581	equip8_3
0,85328	**equip32_2**
0,847534	equip35_2
0,84403	equip2_1
0,843003	equip24_3
0,833353	equip19_3
0,832945	equip15_1
0,826735	equip18_3
0,818713	equip15_2
0,80722	equip33_1
0,77312	equip12_2
0,693537	**equip27_1**
0,692067	equip20_3
0,688618	equip25_2
0,635662	**equip32_1**
0,624787	equip7_3

Quadro 4.11 – Hastes de extensômetros e suas comunalidades – Grupo 2.

Este tipo de análise não foi encontrada na literatura, destacando-se como uma das contribuições inéditas deste trabalho. Além disso, recomenda-se que esta análise (processo de hierarquização) seja repetida periodicamente (a critério dos especialistas na área – engenheiros de Itaipu – que poderia ser, por exemplo, a cada 2 anos), de acordo com a necessidade indicada por especialistas da área. Isto pode mostrar o surgimento de novas hastes indicadas para intensificação de leituras (que deveriam ser investigadas), o mesmo ocorrendo com hastes que poderiam deixar de ser indicadas.

Comunalidade	Haste
0,975487	**equip29_1**
0,966626	equip23_1
0,952144	equip23_2
0,946772	equip24_1
0,945604	**equip22_1**
0,943369	equip24_2
0,942178	**equip34_1**
0,928596	**equip22_2**
0,924851	**equip22_3**
0,917463	equip19_2
0,909212	equip11_1
0,899795	equip14_1
0,866861	equip28_2
0,853862	equip13_1
0,845538	equip25_1
0,812306	equip20_1
0,783119	equip7_1
0,766859	**equip34_2**
0,732343	equip7_2
0,730472	equip18_1
0,680493	equip18_2
0,660694	equip28_1
0,647952	equip19_1
0,60656	equip8_1

Quadro 4.12 – Hastes de extensômetros e suas comunalidades – Grupo 3.

Quanto há hastes dentro dos grupos com comunalidades baixas, recomenda-se a investigação das mesmas. A comunalidade baixa indica alta porcentagem de aleatoriedade nos dados e isto pode ser um indicador de problemas com a haste.

Esta identificação de hastes semelhantes também pode ser útil na projeção de valores de controle. Neste caso, os valores de controle para cada haste podem estar associados às leituras das hastes pertencentes ao mesmo grupo.

Já o escore fatorial final faz a hierarquização dos atributos sendo que, neste caso, os padrões são vetores cujas componentes (atributos) são as leituras das hastes dos extensômetros num determinado mês. Portanto, o escore fatorial final faz a hierarquização dos meses, mostrando se há algum mês com maior relevância, merecedor de maior atenção.

O quadro 4.13 mostra os 15 primeiros meses com maior escore fatorial final e os 15 últimos meses com menor escore fatorial final, considerando as 72 hastes de extensômetros. Os valores dos 15 primeiros meses com maior escore fatorial final revelam que todos os meses do ano são importantes, não há um ou mais meses relevantes. Somente o mês de dezembro não aparece entre os 15 primeiros

meses. Observou-se que o ano mais relevante foi o ano de 1995 e, analisando-se a temperatura ambiente no período estudado, verificou-se que isto ocorreu devido à alta variação na temperatura. Os valores dos 15 últimos meses com menor escore fatorial final revelam que os meses abril, maio e junho são os mais importantes, identificando o efeito do verão.

15 primeiros		15 últimos	
escore fatorial final	**Mês**	**escore fatorial final**	**Mês**
1,755	Janeiro/98	-0,761	Abril/02
1,217	Agosto/95	-0,816	Julho/03
1,153	Janeiro/95	-0,821	Fevereiro/03
1,091	Fevereiro/95	-0,821	Junho/00
0,992	Junho/95	-0,856	Junho/02
0,914	Março/95	-0,877	Maio/03
0,902	Abril/95	-0,904	Fevereiro/00
0,877	Novembro/96	-0,924	Maio/00
0,794	Maio/95	-0,934	Maio/02
0,781	Julho/95	-0,965	Abril/00
0,776	Abril/97	-0,971	Abril/04
0,749	Outubro/95	-1,050	Abril/03
0,741	Novembro/95	-1,061	Junho/03
0,710	Setembro/95	-1,135	Maio/00
0,709	Fevereiro/96	-1,152	Março/03

Quadro 4.13 – Escore fatorial final dos meses de leitura das 72 hastes de extensômetros.

Como foi mostrado na figura 4.2, o grupo 1 apresenta o efeito verão/inverno em suas leituras. Por este motivo, calculou-se o escore fatorial final para fazer a hierarquização dos meses para o grupo 1, para mostrar se há algum mês ou alguns meses com maior relevância.

O quadro 4.14 mostra os 15 primeiros meses com maior escore fatorial final e os 15 últimos meses com menor escore fatorial final, considerando as 11 hastes de extensômetros do grupo1. Os valores dos 15 primeiros meses com maior escore fatorial final revelam que os meses de setembro, outubro e novembro são os mais relevantes, identificando o efeito do inverno. Os valores dos 15 últimos meses com menor escore fatorial final revelam que os meses março, abril, maio e junho são os mais importantes, identificando o efeito do verão.

Este tipo de análise também não foi encontrada na literatura. A identificação de meses com leituras mais significativas para um efeito externo (neste caso, o efeito do verão e do inverno nas leituras das hastes de extensômetros), pode ser útil na projeção de valores de controle, por exemplo. Admitindo-se que há distinção nas

leituras das hastes para estes meses, somente leituras realizadas nestes meses seriam utilizadas para definir valores de controle específicos para estes meses.

15 primeiros		15 últimos	
escore fatorial final	**Mês**	**escore fatorial final**	**Mês**
1,752	Outubro/00	-0,896	Agosto/96
1,749	Outubro/04	-0,911	Abril/98
1,730	Setembro/04	-0,914	Junho/99
1,664	Outubro/01	-0,918	Outubro/97
1,652	Setembro/00	-0,919	Maio/98
1,647	Novembro/04	-0,929	Março/00
1,632	Setembro/01	-0,932	Maio/96
1,602	Novembro/00	-0,945	Junho/97
1,569	Dezembro/04	-0,956	Abril/96
1,548	Novembro/01	-0,964	Junho/96
1,504	Agosto/04	-0,966	Março/98
1,494	Agosto/01	-0,986	Março/99
1,398	Setembro/02	-0,995	Maio/99
1,394	Outubro/02	-0,999	Abril/99
1,322	Setembro/03	-1,086	Julho/96

Quadro 4.14 – Escore fatorial final dos meses de leitura das 11 hastes de extensômetros – Grupo 1.

4.2 APLICAÇÃO DAS REDES NEURAIS DE KOHONEN UNIDIMENSIONAL PARA O AGRUPAMENTO

As Redes Neurais de Kohonen Unidimensional (1D-SOM) foram aplicadas às cinco bases de dados (IRIS, WINE, PIMA, GUN e LIGHTNING-2), onde era conhecido o grupo a que cada padrão pertence. Por não se tratar de um método exato, ou seja, há variação nos resultados quando aplicado por diversas vezes, este método foi aplicado a cada base de dados por 10 vezes. Os resultados da aplicação dos métodos apresentados nos quadro 4.15, a seguir, consistem nas médias para as 10 vezes.

BASES	**R** (quanto maior, melhor)	**F** (quanto maior, melhor)	**Classificação Errada** (%)
IRIS	0,863	0,865	12,800
WINE	0,764	0,761	22,416
PIMA	0,549	0,655	34,570
GUN	0,504	0,530	47,100
LIGHTNING-2	0,571	0,678	31,901

Quadro 4.15 – Resultados da aplicação das Redes Neurais de Kohonen Unidimensional para o agrupamento, médias da execução de 10 vezes, para a base de dados IRIS, WINE, PIMA, GUN e LIGHTNING-2.

4.3 RESULTADOS DO ALGORITMO DE AGRUPAMENTO BASEADO EM FORMIGAS PROPOSTO

O algoritmo de Agrupamento baseado em Formigas proposto foi aplicado inicialmente às cinco bases de dados (conhecido o grupo a que cada padrão pertence). Por não se tratar de um método exato, ou seja, há variação nos resultados quando aplicado por diversas vezes, este método também foi aplicado a cada base de dados por 10 vezes.

Para avaliação dos resultados foram utilizadas as seguintes medidas de avaliação do agrupamento: Similaridade (*Sim*), Índice Aleatório (*R*), Medida *F* e percentual de classificação errada. Resultados preliminares para as bases de dados IRIS e WINE foram publicados em Villwock e Steiner (2009a, 2009b).

4.3.1 Resultados da Aplicação do Algoritmo de Agrupamento Baseado em Formigas Proposto para as 5 Bases de Dados

O quadro 4.16, a seguir, apresenta a média e o desvio-padrão das medidas de avaliação para as bases de dados reais. Este quadro também apresenta as medidas de avaliação do agrupamento para o melhor resultado.

Resultados		**R**	**F**	**Classificação errada (%)**
IRIS	**Média**	0,871	0,877	11,9
	Desvio-padrão	0,039	0,050	4,6
	Melhor resultado	0,927	0,940	6,0
WINE	**Média**	0,843	0,871	12,7
	Desvio-padrão	0,019	0,021	1,9
	Melhor resultado	0,871	0,899	10,1
PIMA	**Média**	0,510	0,583	43,6
	Desvio-padrão	0,010	0,022	4,0
	Melhor resultado	0,531	0,623	37,5

Quadro 4.16 – Resultados da aplicação do algoritmo de Agrupamento baseado em Formigas proposto, médias da execução de 10 vezes, para as bases de dados reais (IRIS, WINE e PIMA).

As figuras 4.6 e 4.7, a seguir, apresentam a grade para o melhor resultado (cujas medidas de avaliação foram apresentadas no quadro 4.16) para as bases de dados reais (IRIS e WINE), respectivamente. Nestas figuras, os padrões em vermelho pertencem ao grupo 1, os padrões em preto pertencem ao grupo 2 e os padrões em azul pertencem ao grupo 3.

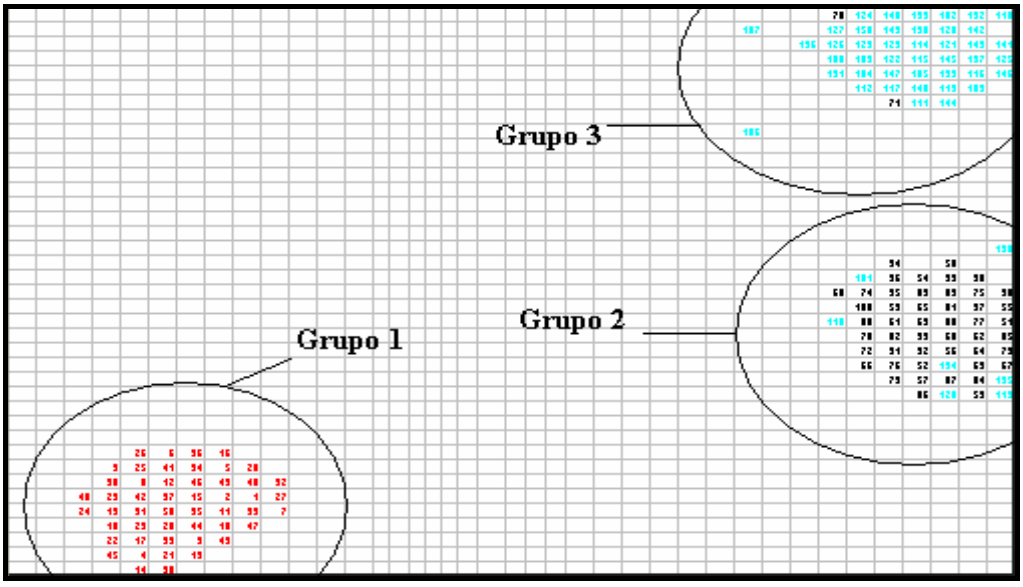

Figura 4.6 – Resultado do algoritmo de agrupamento baseado em Formigas proposto para a base de dados IRIS – melhor resultado.

O quadro 4.17 (matriz de confusão) mostra a distribuição dos padrões para a base de dados IRIS, onde pode-se observar os padrões atribuídos aos grupos corretamente e os padrões atribuídos aos grupos erroneamente. Nesta base de dados são apenas nove padrões em grupos errados de um total de 150 padrões. O grupo 1 contém todos os padrões atribuídos a ele.

IRIS	Solução Gerada		
Agrupamento Correto	**Grupo 1**	**Grupo 2**	**Grupo 3**
Classe 1	50	0	0
Classe 2	0	48	2
Classe 3	0	7	43

Quadro 4.17 – Distribuição dos Padrões – IRIS – melhor resultado

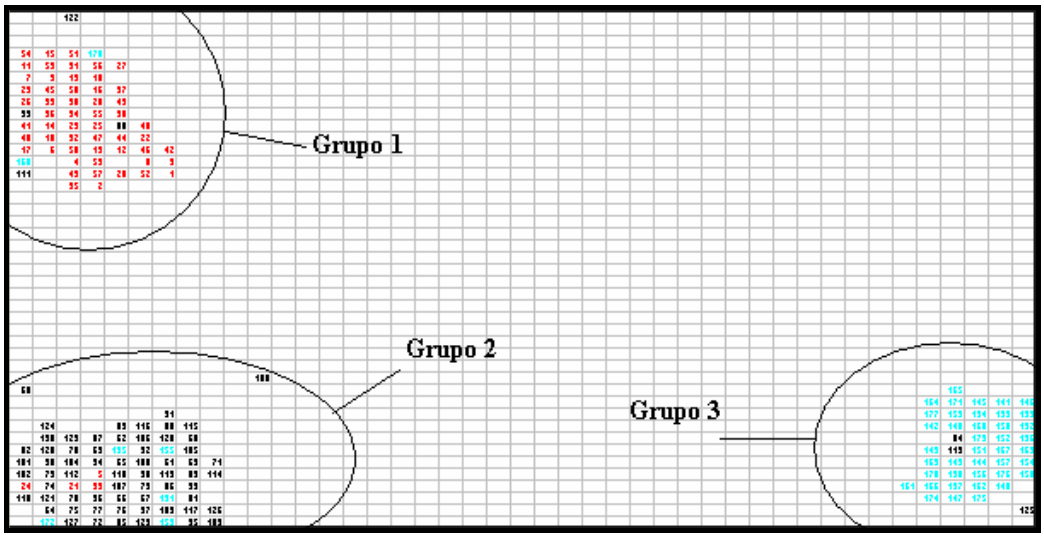

Figura 4.7 – Resultado do algoritmo de agrupamento baseado em Formigas proposto para a base de dados WINE – melhor resultado.

Da mesma forma, o quadro 4.18 mostra a distribuição dos padrões para a base de dados WINE. Nesta base de dados são 18 padrões em grupos errados de um total de 178 padrões.

WINE	Solução Gerada		
Agrupamento Correto	**Grupo 1**	**Grupo 2**	**Grupo 3**
Classe 1	55	4	0
Classe 2	4	64	3
Classe 3	2	5	41

Quadro 4.18 – Distribuição dos Padrões – WINE – melhor resultado

O quadro 4.19, a seguir, apresenta a média e o desvio-padrão das medidas de avaliação para as bases de dados de séries temporais. Este quadro também apresenta as medidas de avaliação do agrupamento para o melhor resultado.

Resultados		**R**	**F**	**Classificação errada (%)**
GUN	**Média**	0,535	0,611	38,2
	Desvio-padrão	0,038	0,063	7,4
	Melhor resultado	0,618	0,729	25,5
LIGHTNING-2	**Média**	0,556	0,647	34,1
	Desvio-padrão	0,038	0,054	7,2
	Melhor resultado	0,608	0,706	26,4

Quadro 4.19 – Resultados da aplicação do algoritmo de Agrupamento baseado em Formigas proposto, médias da execução de 10 vezes, para as bases de dados de séries temporais (GUN e LIGHTNING-2).

As figuras 4.8 e 4.9, a seguir, apresentam a grade para o melhor resultado (cujas medidas de avaliação foram apresentadas no quadro 4.19) das bases de dados GUN e LIGHTNING-2, respectivamente. Nestas bases de dados, os padrões em vermelho pertencem ao grupo 1 e os padrões em preto pertencem ao grupo 2.

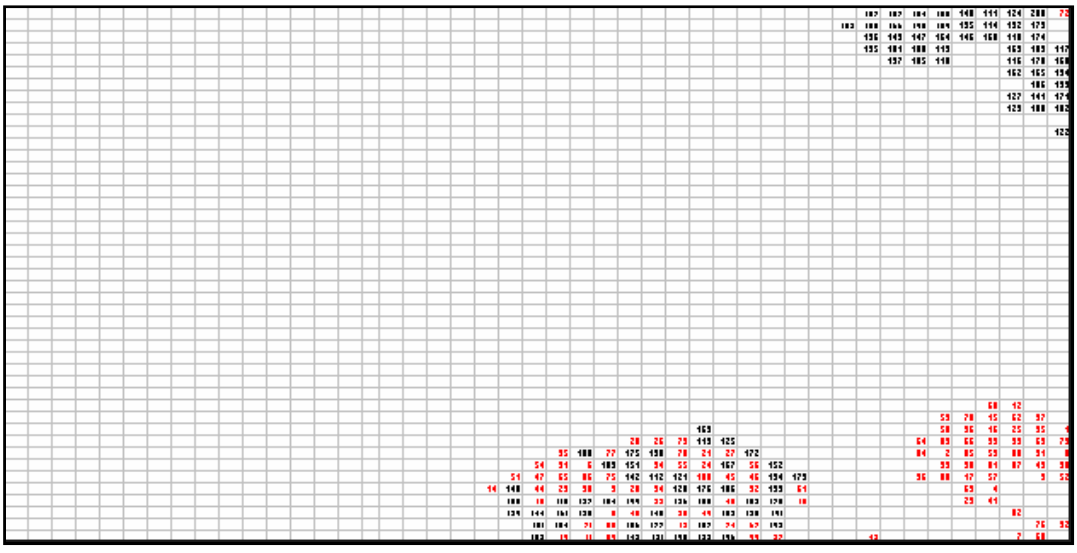

Figura 4.8 – Resultado do algoritmo de agrupamento proposto baseado em Formigas para a base de dados GUN – melhor resultado.

O quadro 4.20 mostra a distribuição dos padrões para a base de dados GUN. Nesta base de dados são 51 padrões em grupos errados de um total de 200 padrões.

GUN	Solução Gerada	
Agrupamento Correto	**Grupo 1**	**Grupo 2**
Classe 1	99	1
Classe 2	50	50

Quadro 4.20 – Distribuição dos Padrões – GUN – melhor resultado

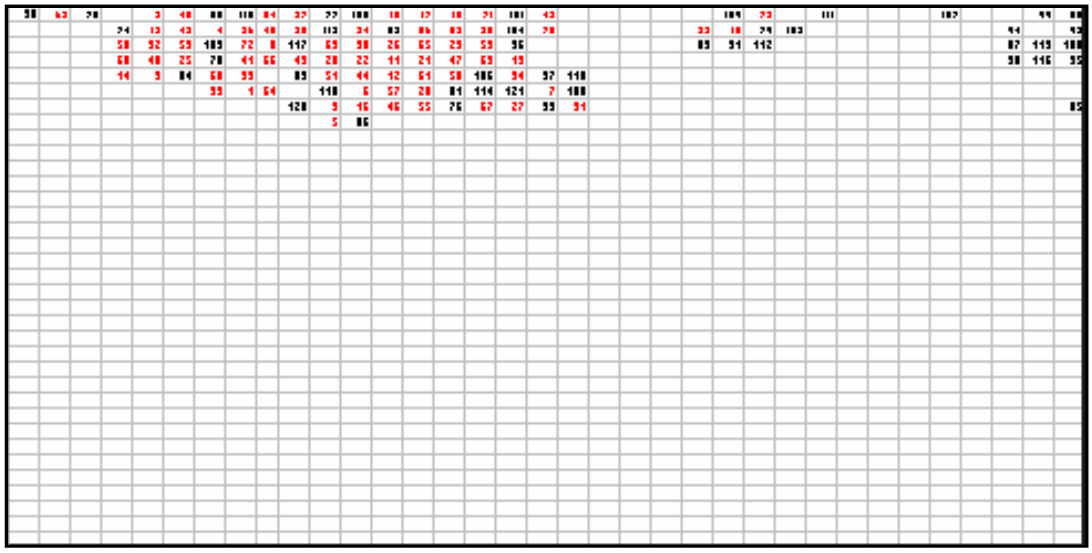

Figura 4.9 – Resultado do algoritmo de agrupamento proposto baseado em Formigas para a base de dados LIGHTNING-2 – melhor resultado.

E, finalmente, o quadro 4.21 mostra a distribuição dos padrões para a base de dados LIGHTNING-2. Nesta base de dados são 32 padrões em grupos errados de um total de 121 padrões.

LIGHTNING-2	Solução Gerada	
Agrupamento Correto	**Grupo 1**	**Grupo 2**
Classe 1	70	3
Classe 2	29	19

Quadro 4.21 – Distribuição dos Padrões – LIGHTNING-2 – melhor resultado

4.3.2 Avaliação do Algoritmo de Agrupamento por Formigas Proposto em relação a outros dois métodos – Método Ward e Redes Neurais de Kohonen Unidimensional

Nos quadros 4.22, 4.23 e 4.24 são apresentadas as comparações das medidas médias de avaliação para os três métodos, para as bases de dados IRIS, WINE e PIMA, respectivamente. O melhor resultado encontra-se em negrito.

IRIS	Ward	1D-SOM	Formigas
R (quanto maior melhor)	**0,957**	0,863	0,871
F (quanto maior melhor)	**0,967**	0,865	0,877
Classificação errada (%) (quanto menor melhor)	**3,333**	12,8	11,9

Quadro 4.22 – Comparação dos resultados médios da aplicação dos métodos de agrupamento Ward, Redes Neurais de Kohonen Unidimensional e Agrupamento baseado em Formigas proposto para a base de dados IRIS.

WINE	Ward	1D-SOM	Formigas
R (quanto maior melhor)	0,819	0,764	**0,843**
F (quanto maior melhor)	0,845	0,761	**0,871**
Classificação errada (%) (quanto menor melhor)	15,169	22,416	**12,7**

Quadro 4.23 – Comparação dos resultados médios da aplicação dos métodos de agrupamento Ward, Redes Neurais de Kohonen Unidimensional e Agrupamento baseado em Formigas proposto para a base de dados WINE.

PIMA	Ward	1D-SOM	Formigas
R (quanto maior melhor)	0,531	**0,549**	0,510
F (quanto maior melhor)	0,624	**0,655**	0,583
Classificação errada (%) (quanto menor melhor)	37,370	**34,570**	43,6

Quadro 4.24 – Comparação dos resultados médios da aplicação dos métodos de agrupamento Ward, Redes Neurais de Kohonen Unidimensional e Agrupamento baseado em Formigas proposto para a base de dados PIMA.

Os resultados não mostram superioridade de algum método. Na base de dados IRIS, o Método Ward foi melhor (cerca de 3% de erros); na base de dados WINE, o Algoritmo baseado em Formigas proposto foi melhor (cerca de 12% de erros) e na base de dados PIMA, as Redes Neurais de Kohonen Unidimensional foi melhor (cerca de 34% de erros).

Handl, Knowles e Dorigo (2006) também afirmam que nenhum algoritmo domina os outros sempre. Segundo Ho e Pepyne (2002), pelo teorema "*NO-FREE-LUNCH*", se não há nenhuma suposição anterior sobre o problema de otimização que se tenta resolver, é de se esperar que nenhuma estratégia tenha melhor desempenho que outra quando testada em um conjunto grande de bases de dados com características diversas.

Nos quadros 4.25 e 4.26, a seguir, são apresentadas as comparações das medidas médias de avaliação para os três métodos, para as bases de dados GUN e LIGHTNING-2, respectivamente. O melhor resultado encontra-se em negrito.

GUN	Ward	1D-SOM	Formigas
R (quanto maior melhor)	0,497	0,504	**0,535**
F (quanto maior melhor)	0,500	0,530	**0,611**
Classificação errada (%) (quanto menor melhor)	50,000	47,100	**38,2**

Quadro 4.25 – Comparação dos resultados médios da aplicação dos métodos de agrupamento Ward, Redes Neurais de Kohonen Unidimensional e Agrupamento baseado em Formigas proposto para a base de dados GUN.

LIGHTNING-2	Ward	1D-SOM	Formigas
R (quanto maior melhor)	**0,633**	0,571	0,556
F (quanto maior melhor)	**0,739**	0,678	0,647
Classificação errada (%) (quanto menor melhor)	**23,967**	31,901	34,1

Quadro 4.26 – Comparação dos resultados médios da aplicação dos métodos de agrupamento Ward, Redes Neurais de Kohonen Unidimensional e Agrupamento baseado em Formigas proposto para a base de dados LIGHTNING-2.

Os resultados não mostram superioridade de algum método. Na base de dados GUN, o Algoritmo baseado em Formigas proposto foi melhor (cerca de 38% de erros) e, na base de dados LIGHTNING-2, o Método Ward foi melhor (cerca de 24% de erros). Os cerca de 38% de erros para a base de dados GUN não é um resultado satisfatório, porém, já é melhor do que nos outros dois métodos.

O quadro 4.27, a seguir, apresenta a comparação das médias das medidas de avaliação do agrupamento para o algoritmo proposto e para o algoritmo *ACAM* (*Ant-based Clustering Algorithm Modified*) proposto por Boryczka (2009). O melhor resultado encontra-se em negrito. Os resultados mostram que o algoritmo proposto é melhor que o ACAM para duas das três bases de dados.

BASES	Medidas de Avaliação	*ACAM*	Algoritmo Proposto
IRIS	**R**	0,819	**0,871**
	F	0,810	**0,877**
	Classificação errada (%)	18,7	**11,9**
WINE	**R**	**0,849**	0,843
	F	0,868	**0,871**
	Classificação errada (%)	13,9	**12,7**
PIMA	**R**	**0,522**	0,510
	F	0,574	**0,583**
	Classificação errada (%)	**33,7**	43,6

Quadro 4.27 – Comparação dos resultados médios da aplicação do algoritmo proposto com resultados disponíveis em Boryczka (2009) para as bases de dados reais.

O quadro 4.28 apresenta a medida de similaridade média (*sim*) para comparação com um resultado disponível em Keogh (2006). As bases de dados disponíveis em Keogh (2006) foram divididas em dois conjuntos (treinamento e teste) e os resultados apresentados referem-se às medidas realizadas no conjunto de teste. Nos resultados apresentados em Keogh (2006), O melhor resultado encontra-se em negrito.

	Sim (quanto maior, melhor)	
BASES	**Keogh (2006)**	**Algoritmo Proposto**
GUN	0.500	**0,598**
LIGHTNING-2	**0.611**	0,610

Quadro 4.28 – Comparação dos resultados médios da aplicação do algoritmo proposto com resultados disponíveis em Keogh (2006) para as bases de dados de séries temporais.

Aqui, também, os resultados não mostram superioridade de algum método. Isto mostra que a investigação deve ser ampliada, mas não inviabiliza o uso do algoritmo proposto em bases de dados de séries temporais.

De qualquer forma, estes resultados qualificam o algoritmo proposto para aplicação aos dados de instrumentação geotécnica-estrutural da Usina Hidrelétrica de Itaipu.

4.3.3 Resultados da Aplicação do Algoritmo de Agrupamento Baseado em Formigas Proposto para os Dados de Instrumentação Geotécnica-estrutural da Itaipu

O quadro 4.29, a seguir, apresenta os resultados médios desta aplicação (instrumentação geotécnica-estrutural da Itaipu), com respeito as medidas de avaliação variância e Índice *Dunn*. Estas medidas foram utilizadas, já que para estes dados não havia conhecimento prévio do grupo a que cada padrão pertencia.

Resultados		**Variância**	**D**
Itaipu	**Média**	2,800	2,397
	Desvio-padrão	1,227	1,367
	Melhor resultado	1,231	5,547

Quadro 4.29 – Resultados da avaliação do agrupamento pelo algoritmo de Agrupamento baseado em Formigas proposto para os dados de instrumentação geotécnica-estrutural da Itaipu.

Estas medidas foram calculadas levando-se em consideração que os dados devem ser agrupados em três grupos, conforme discussão apresentada na seção 4.1. Na aplicação do algoritmo proposto, em apenas uma das 10 vezes que o algoritmo foi aplicado, houve a identificação visual de três grupos. Em outras seis aplicações, apenas dois grupos são visualizados.

O melhor resultado desta aplicação, segundo a variância, é apresentado na figura 4.10. Apesar deste resultado apresentar somente dois grupos visualmente, o resultado pode ser considerado satisfatório, pois conseguiu identificar os 11 instrumentos pertencentes ao grupo 1 (também identificados pelo Método Ward), considerado o grupo das hastes de maior importância. Os padrões circulados em vermelho, nesta figura 4.10, indica que os mesmos não deveriam ter sido atribuídos ao grupo 1.

Handl, Knowles e Dorigo (2006) afirmam que seu algoritmo, freqüentemente, não identifica o número correto de grupos. Os resultados indicaram que os grupos identificados por eles, correspondem a estrutura dos dados e que estes dados, por sua vez, possuem estruturas de agrupamento em vários níveis. Quando as várias classes nos dados não são bem separadas, também não serão espacialmente separáveis, mas podem ser observadas estruturas de agrupamento em um nível mais grosseiro. Isto também foi observado neste estudo.

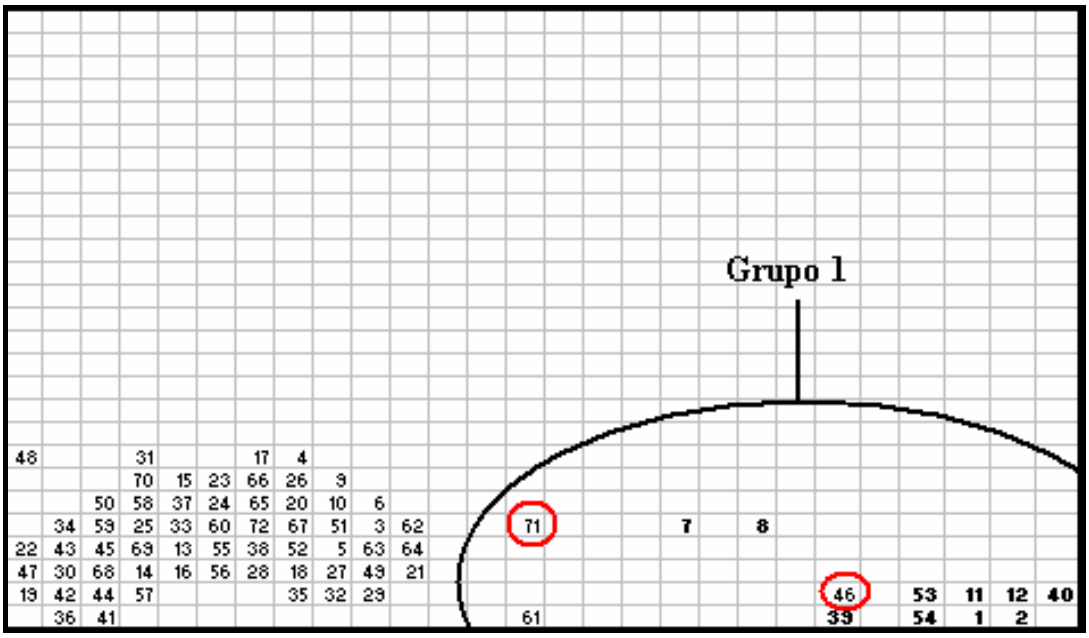

Figura 4.10 – Resultado do algoritmo de Agrupamento baseado em Formigas proposto para os dados de instrumentação geotécnica-estrutural da Barragem de Itaipu – melhor resultado.

As figuras 4.11 e 4.12 mostram o resultado onde houve a identificação visual de três grupos e o melhor resultado, respectivamente, colorindo-se os padrões conforme resultado obtido pelo Método Ward. Os padrões em vermelho pertencem ao grupo 1, os padrões em preto pertencem ao grupo 2 e os padrões em azul pertencem ao grupo 3. Observa-se que o algoritmo proposto gera praticamente os mesmos grupos, mesmo onde a identificação visual aponta para dois grupos.

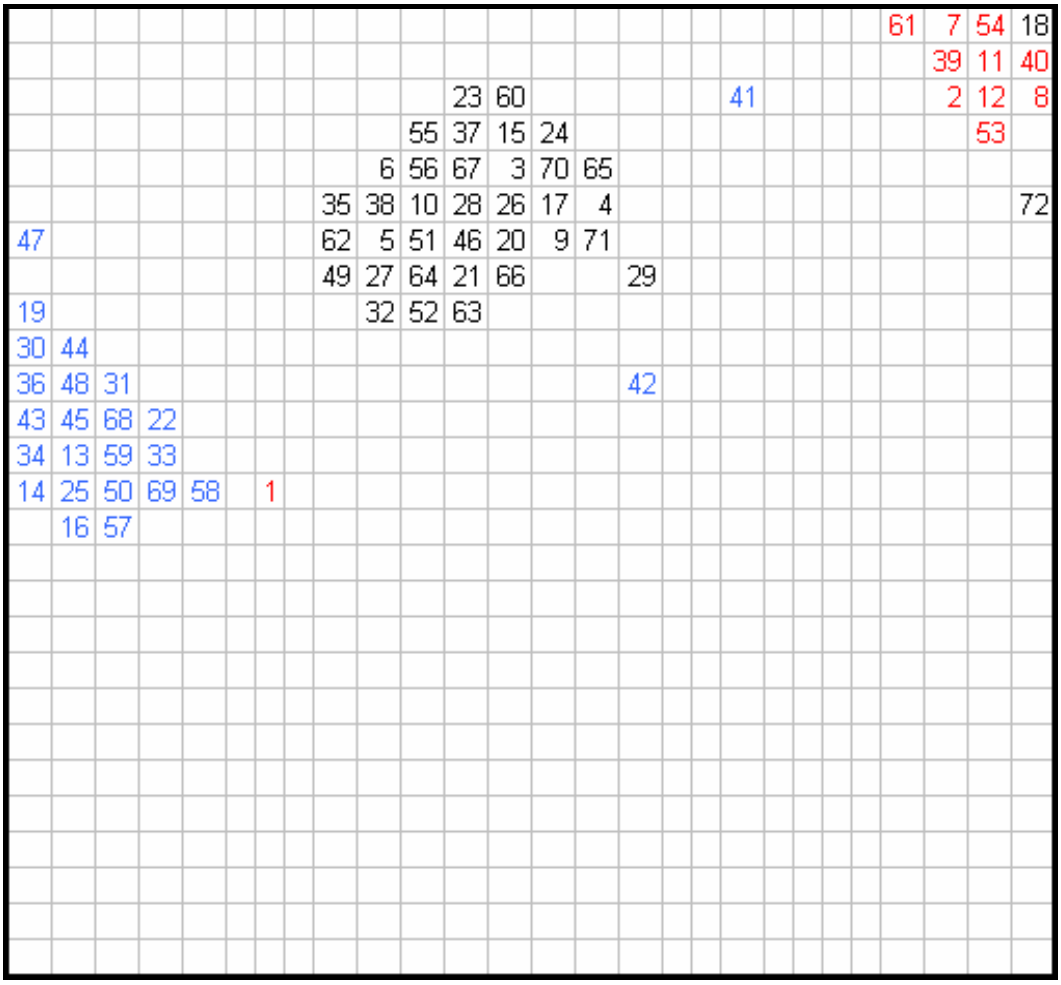

Figura 4.11 – Resultado do algoritmo de Agrupamento baseado em Formigas proposto para os dados de instrumentação geotécnica-estrutural da Barragem de Itaipu – resultado com identificação visual de 3 grupos.

No resultado da figura 4.12, o algoritmo proposto deixa os padrões dos grupos 2 e 3 visualmente em um único grupo, mas separados (como pode ser observado em função das cores). O objetivo de uma das modificações propostas era justamente tentar resolver isso. Porém, como isso continuou acontecendo, acredita-se que estes grupos não são bem separados e, neste caso, pode ser observada somente uma estrutura de agrupamento em um nível mais grosseiro, como explicado anteriormente.

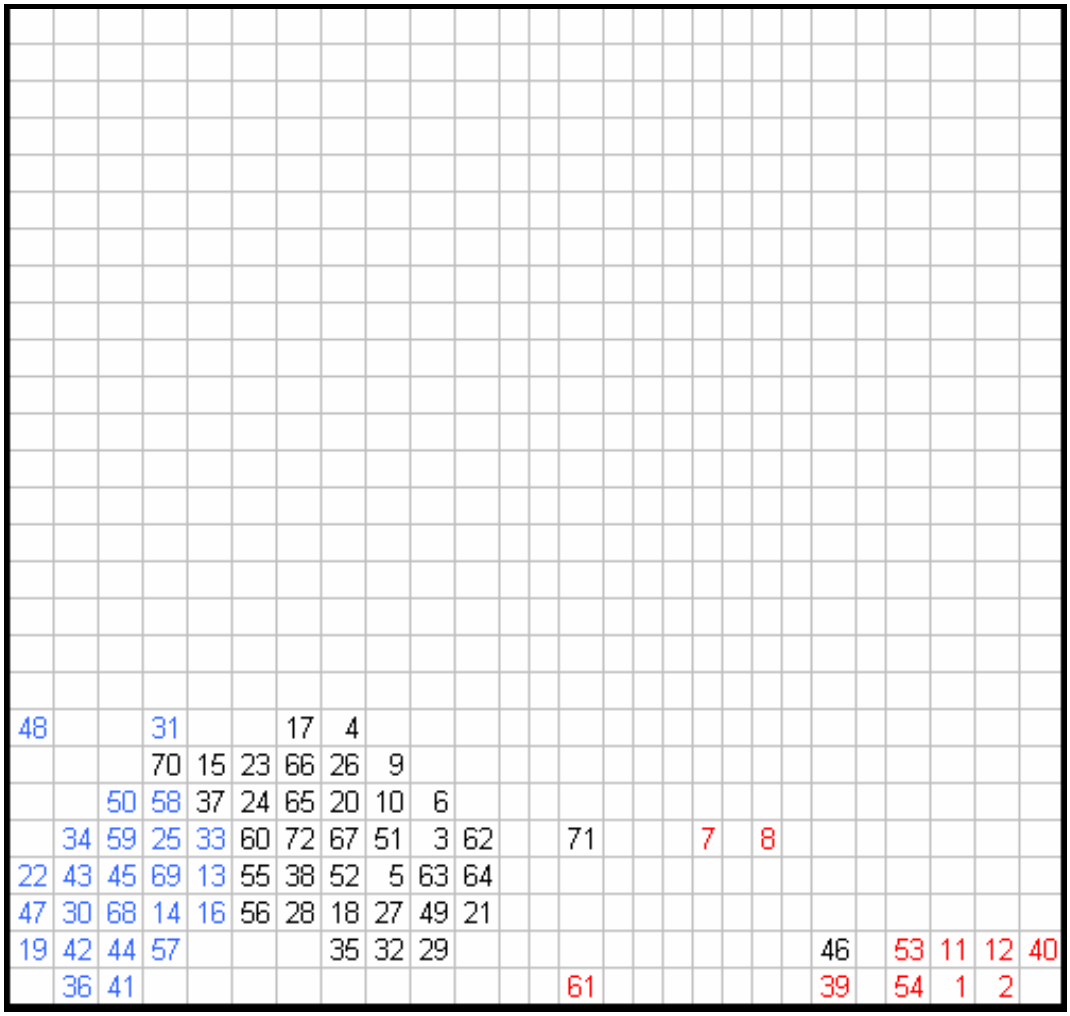

Figura 4.12 – Resultado do algoritmo de Agrupamento baseado em Formigas proposto para os dados de instrumentação geotécnica-estrutural da Barragem de Itaipu – melhor resultado – comparação com o Método Ward.

No quadro 4.30 são apresentadas as comparações das variâncias médias para os três métodos, para os dados de instrumentação geotécnica-estrutural da Itaipu. Em todos os métodos foi considerado o número de grupos igual a “3”. O melhor resultado encontra-se em negrito.

Itaipu	Ward	1D-SOM	Formigas
variância (quanto menor melhor)	0,782	**0,312**	2,800

Quadro 4.30 – Comparação das variâncias médias da aplicação dos métodos de agrupamento Ward, Redes Neurais de Kohonen Unidimensional e Agrupamento proposto baseado em Formigas para os dados de instrumentação geoténica-estrutural da Itaipu.

Nos dados de instrumentação geotécnica-estrutural da Itaipu, o agrupamento através das Redes Neurais de Kohonen Unidimensional foi melhor. O

quadro 4.31 mostra a distribuição das hastes em cada grupo para o melhor resultado conforme apresentado no quadro 4.32. Porém, a visualização do agrupamento em duas dimensões promovido pelo Agrupamento baseado em Formigas, justifica sua aplicação. Além disso, este resultado era esperado visto que no método Redes Neurais de Kohonen Unidimensional, o número de neurônios igual a três, implica na formação de três grupos. Esta pode ser considerada uma vantagem deste método em relação aos demais.

Grupo 1	Grupo 2	Grupo 3
equip1_1	equip2_2	equip2_1
equip1_2	equip3_1	equip7_1
equip4_1	equip3_2	equip7_2
equip4_2	equip5_1	equip7_3
equip6_1	equip5_2	equip8_1
equip6_2	equip8_2	equip11_1
equip8_3	equip12_1	equip13_1
equip15_2	equip12_2	equip13_2
equip18_3	equip14_2	equip13_3
equip21_1	equip14_3	equip14_1
equip21_2	equip15_1	equip18_1
equip26_1	equip19_3	equip18_2
equip26_2	equip20_3	equip19_1
equip31_1	equip23_3	equip19_2
equip32_3	equip24_3	equip20_1
equip35_1	equip25_2	equip20_2
equip35_2	equip25_3	equip22_1
	equip32_1	equip22_2
	equip32_2	equip22_3
	equip33_3	equip23_1
		equip23_2
		equip24_1
		equip24_2
		equip25_1
		equip27_1
		equip27_2
		equip28_1
		equip28_2
		equip29_1
		equip29_2
		equip33_1
		equip33_2
		equip34_1
		equip34_2
		equip34_3

Quadro 4.31 – Resultado da aplicação do método de agrupamento Redes Neurais de Kohonen Unidimensional para os dados de instrumentação geoténica-estrutural da Itaipu – melhor resultado.

Resultados		Variância
Itaipu	**Média**	0,312
	Desvio-padrão	0,194
	Melhor resultado	0,034

Quadro 4.32 – Resultados da avaliação do agrupamento pela aplicação das Redes Neurais de Kohonen Unidimensional para os dados de instrumentação geotécnica-estrutural da Itaipu.

5 CONCLUSÕES E SUGESTÕES PARA TRABALHOS FUTUROS

5.1 CONCLUSÕES

Neste trabalho é apresentada uma metodologia que pode ser enquadrada na área de *KDD*, cujo objetivo é selecionar, agrupar e hierarquizar instrumentos geotécnico-estruturais de uma usina hidrelétrica, neste trabalho a Usina Hidrelétrica de Itaipu, maximizando a eficácia e a eficiência das análises das leituras.

A metodologia apresentada foi aplicada aos instrumentos chamados extensômetros, localizados em diferentes pontos do bloco F da barragem, num total de 30 que, com uma, duas ou três hastes, totalizam 72 medidas de deslocamento mensais, armazenadas no decorrer de 10 anos, totalizando 120 leituras (janeiro/1995 a dezembro/2004). Vale relembrar que das 72 medidas, 24 foram automatizadas pela empresa. A hierarquização dos instrumentos seria uma forma de fazer esta escolha sem nenhum conhecimento prévio sobre a localização, feição, ou outra característica dos instrumentos. Desta forma, pode-se pensar em aplicar esta metodologia para novas tomadas de decisão quanto a automatização de instrumentos adicionais.

A metodologia para a análise do problema de Itaipu (ilustrada pelo fluxograma da figura 5.1) ficou composta da seguinte forma:

Na 1ª etapa do processo *KDD*, (seleção dos dados), ficou definido que a metodologia seria aplicada somente aos extensômetros localizados no trecho F, conforme já especificado acima. Na 2ª etapa (pré-processamento dos dados), os dados disponibilizados pela Itaipu foram convertidos para planilhas, das quais foram extraídas as informações necessárias para o desenvolvimento deste trabalho. Na 3ª etapa (formatação dos dados), para a sua posterior aplicação aos métodos de agrupamento (etapa de Mineração de Dados), os dados foram padronizados.

Na 4ª etapa do processo *KDD* (Mineração de Dados), a tarefa realizada foi o de agrupamento de padrões. Nesta etapa foram aplicadas, paralelamente (figura 5.1), a Análise Fatorial (para hierarquizar as 72 hastes de extensômetros) e a Análise de Agrupamento, pelo Método Ward, para agrupar as 72 hastes de extensômetros semelhantes. A Análise Fatorial foi aplicada, também, dentro de cada grupo formado pela Análise de Agrupamento.

Paralelamente, o algoritmo de Agrupamento proposto baseado em Formigas foi aplicado inicialmente às cinco bases de dados. Para avaliação do desempenho do algoritmo proposto, este foi comparado ao Método Ward e as Redes Neurais de Kohonen Unidimensional, utilizando as seguintes medidas de avaliação do agrupamento: Similaridade (*Sim*), Índice Aleatório (*R*), Medida *F* e percentual de classificação errada. Só então, o algoritmo proposto foi aplicado aos dados de instrumentação geotécnica-estrutural da Itaipu.

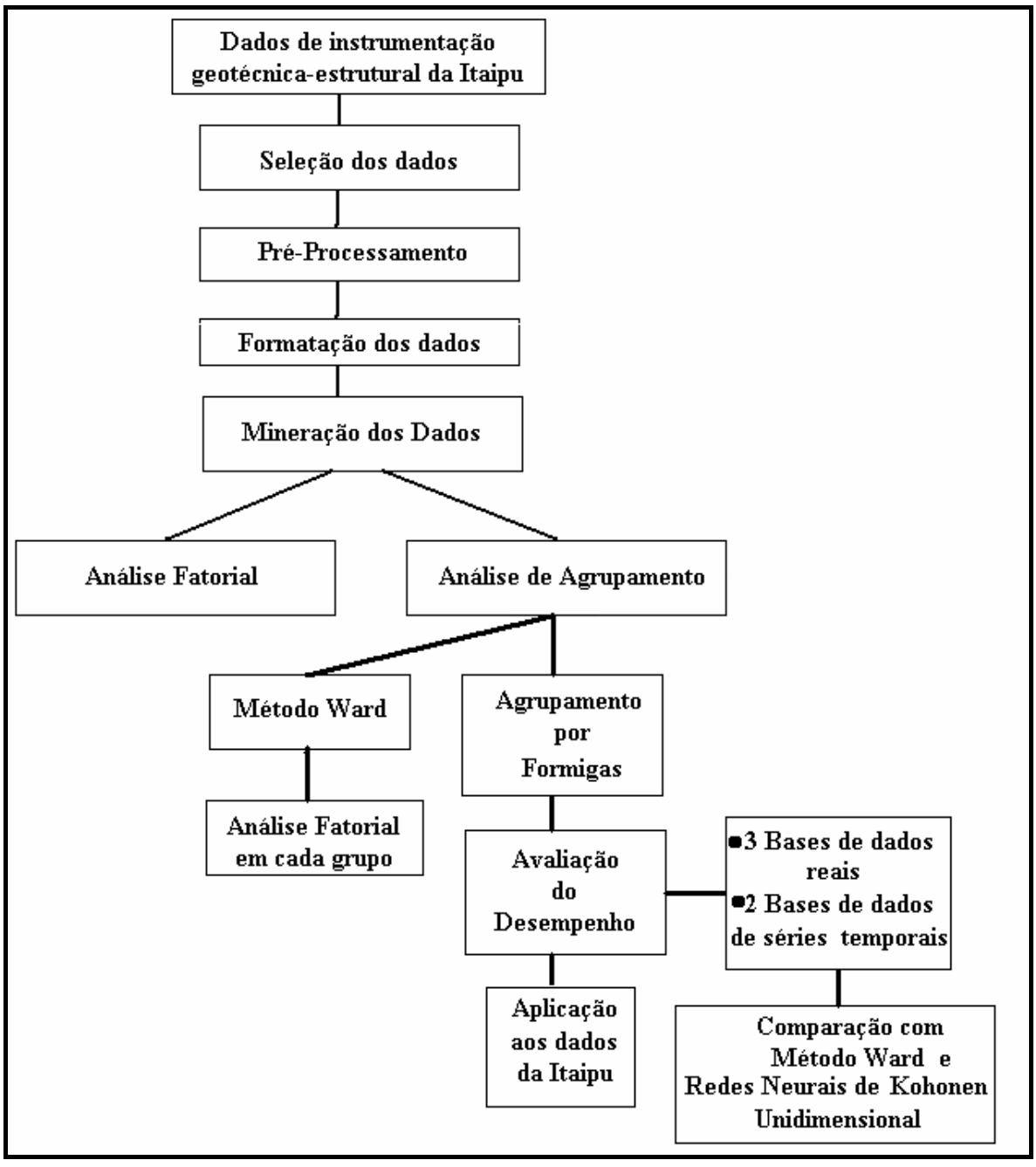

Figura 5.1 – Fluxograma da metodologia empregada neste trabalho.

Na Análise Fatorial aplicada às 72 hastes, não foi detectada necessidade de investigação a nenhuma haste, pois para todas elas, a comunalidade foi alta. No quadro 4.9 são apresentadas as 25 hastes de extensômetros com as comunalidades mais altas. Identificadas em negrito estão as 14 hastes que foram automatizadas pela equipe de Engenheiros da Itaipu (as hastes automatizadas são as consideradas como as mais importantes). O método de hierarquização proposto (sem o prévio agrupamento das hastes) identificou 14 das 24 hastes automatizadas.

Na Análise de Agrupamento, mostrou-se que é possível encontrar justificativas técnicas para a formação dos grupos (quadro 4.6). Os instrumentos foram agrupados segundo características geológicas relevantes do maciço de fundação, apesar das mesmas não terem sido explicitamente apresentadas aos métodos de Mineração de Dados.

Observados os três grupos 1, 2 e 3, aplicou-se a Análise Fatorial dentro de cada grupo para fazer a hierarquização das hastes de extensômetros. Nos quadros 4.9, 4.10 e 4.11, as hastes de extensômetros identificadas em negrito são as automatizadas pela Itaipu. Observou-se que as hastes de extensômetros automatizados estão, na maioria das vezes, entre as primeiras do *ranking* em cada grupo.

Ainda, com o intuito de identificar as 24 hastes mais relevantes, optou-se por identificar as oito hastes melhores ranqueadas em cada grupo. Neste caso, obter-se-ia 15 das 24 hastes automatizadas. Este número de hastes coincidentes com as automatizadas pela Itaipu aumentaria com o auxílio de um especialista para uma melhor interpretação dos resultados. Este especialista detectaria que o grupo 1, por exemplo, é formado por hastes extremamente importantes no monitoramento da barragem e que todas as hastes deste grupo deveriam ser automatizadas.

Abordagens similares a esta podem ser usadas em muitos outros casos, pois em nosso país, existem milhares de grandes obras de Engenharia Civil que contam com sistemas de instrumentação, cujos dados podem e devem receber um tratamento adequado.

Paralelamente a Análise Multivariada aplicada aos dados de instrumentação geotécnica-estrutural da Itaipu, o algoritmo de Agrupamento proposto baseado em Formigas foi aplicado inicialmente às cinco bases de dados (bases nas quais conhecia-se, previamente, o grupo a que cada padrão pertence).

Quando comparadas as médias das medidas de avaliação (quadros 4.22 ao 4.26), na aplicação dos métodos de agrupamento Ward, Redes Neurais de Kohonen Unidimensional e Agrupamento proposto baseado em Formigas, para as cinco bases de dados, os resultados não mostraram superioridade de algum dos métodos. Handl, Knowles e Dorigo (2006) também afirmam que nenhum algoritmo domina os outros sempre.

Já na comparação das médias das medidas de avaliação (quadro 4.27) do agrupamento para o algoritmo proposto e para o algoritmo *ACAM* (*Ant-based Clustering Algorithm Modified*), proposto por Boryczka (2009), os resultados mostram que o algoritmo proposto apresentou um desempenho melhor do que o *ACAM* para duas das três bases de dados.

Na comparação das médias (quadro 4.28) da medida de similaridade (*sim*) do agrupamento para o algoritmo proposto com o resultado disponível em Keogh (2006), os resultados não mostram superioridade de algum método. Isto mostra que a investigação deve ser ampliada.

Deste modo, estes resultados qualificam o algoritmo proposto para aplicação aos dados de instrumentação geotécnica-estrutural da Itaipu.

No melhor resultado da aplicação do algoritmo proposto aos dados de instrumentação geotécnica-estrutural da Itaipu, segundo a variância, somente dois grupos são visualmente observados (figura 4.10). Este resultado pode ser considerado satisfatório, pois foi possível identificar os 11 instrumentos pertencentes ao grupo 1 (também identificados pelo Método Ward), considerado o grupo das hastes de maior importância.

Além disso, a aplicação dos métodos de agrupamento às bases de dados de séries temporais sem nenhum método de pré-processamento dos dados visando o agrupamento especificamente para séries temporais, mostrou-se satisfatório.

5.2 PRINCIPAIS CONTRIBUIÇÕES DO TRABALHO

Este trabalho apresenta três principais contribuições, dentre outras consideradas secundárias. Na abordagem de um importante problema de engenharia, a análise de dados de instrumentação de grandes obras, foram aplicadas técnicas de agrupamento, dentre outras, no contexto de *KDD,* do inglês *“Knowledge Discovery in Databases”* ou “Descoberta de Conhecimento em Bases de

Dados", tendo como objetivo a identificação dos instrumentos que são realmente significativos à análise do comportamento de uma barragem.

As novas propostas apresentadas ao algoritmo de Agrupamento baseado em Colônia de Formigas formam a segunda grande contribuição deste trabalho. Esta metaheurística, relativamente nova, ainda exige muita investigação para melhorar seu desempenho.

Durante o estudo do Agrupamento baseado em Formigas, foi observado que muitas das mudanças de posição dos padrões ocorrem desnecessariamente. Considera-se uma mudança desnecessária quando um padrão está entre similares na grade e, neste caso, não há necessidade da mudança deste padrão para outra posição. Com o objetivo de evitar estas mudanças desnecessárias, uma comparação da probabilidade de descarregar um padrão na posição escolhida aleatoriamente com a probabilidade de descarregar este padrão em sua posição atual foi introduzida.

Também foi observada a ocorrência de fusão de grupos próximos na grade. Quando a decisão de descarregar um padrão for positiva e a célula em que o padrão deveria ser descarregado está ocupada, busca-se aleatoriamente uma posição vizinha a esta, que esteja livre. Porém, esta nova posição pode estar próxima também a outro grupo de padrões na grade. Este pode ser um motivo para a fusão de grupos próximos. Como uma alternativa para evitar a fusão de grupos próximos na grade, foi proposta neste trabalho uma avaliação da probabilidade para a nova posição.

Outra questão observada no Agrupamento baseado em Formigas é que uma formiga pode carregar um padrão que está entre similares na grade. Uma formiga só carrega um padrão quando este não está entre similares na grade, porém, desde que a formiga carrega um padrão até ela ser sorteada para tentar descarregar o padrão, mudanças ocorrem na vizinhança deste, podendo deixá-lo então entre similares. Sendo assim, esta formiga fica inativa, pois a operação de descarregar o padrão não é executada. Neste caso, foi proposta a substituição do padrão carregado por uma formiga.

A terceira contribuição foi a aplicação deste algoritmo proposto, a bases de dados de séries temporais. Poucos algoritmos de agrupamentos, recentemente criados, têm sido utilizados no agrupamento de séries temporais.

5.3 SUGESTÕES PARA TRABALHOS FUTUROS

Como sugestão para trabalhos futuros, na aplicação do Algoritmo de Agrupamento proposto baseado em Formigas, recomenda-se o estudo da utilização de outra forma de definição da função *f* como, por exemplo, contando o número de vizinhos similares e o número de vizinhos dissimilares.

Além disso, recomenda-se testes de outras medidas de similaridades ou dissimilaridades, principalmente na aplicação a bases de dados de séries temporais. Outro teste possível é a aplicação dos métodos sem a padronização dos dados. Sugerem-se ainda a utilização de bases de dados adicionais para os testes, principalmente bases de dados de séries temporais, bem como a utilização de mais índices de avaliação do agrupamento.

Também é sugerido o uso de outros critérios de parada como, por exemplo, o uso da quantidade de formigas inativas no processo, introduzida neste trabalho. Além disso, o uso de pesos na grade e o uso de múltiplas colônias de formigas, vistos na revisão de literatura, também podem ser promissores.

A comparação do algoritmo proposto com outros métodos de agrupamento e mapeamento topográfico, também é sugerida para trabalhos futuros. Uma sugestão é o uso das Redes Neurais de Kohonen Bidimensional (2D-SOM). Na aplicação das Redes Neurais de Kohonen Unidimensional (1D-SOM), sugere-se a aplicação do método usando-se um número de neurônios maior que o número de grupos, com o objetivo de “retirar” a vantagem da aplicação do método, de informar o número de grupos que deve ser formado, como observado na seção 4.3.3.

Na aplicação dos métodos de agrupamento às bases de dados de séries temporais, não foi aplicado nenhum método de pré-processamento dos dados visando o agrupamento dos dados especificamente para séries temporais. Sugere-se a aplicação de um ou mais métodos de pré-processamento para avaliação da qualidade do agrupamento, comparando-se o ganho na qualidade com o esforço computacional exigido neste processo.

Na aplicação do processo *KDD* para os dados de instrumentação geotécnica-estrutural da Itaipu, sugere-se a repetição do processo para outros instrumentos, outros períodos e implementação do processo para definição de valores de controle e detecção de anomalias. O processo de hierarquização repetido em diversos períodos (a cada 2 anos, por exemplo) pode mostrar o surgimento de

novas hastes indicadas para intensificação de leituras ou hastes que poderiam deixar de ser indicadas (que deveriam ser investigadas).

Por não se tratar de um método exato, ou seja, há variação nos resultados quando aplicado por diversas vezes, os métodos de Agrupamento por Formigas e Redes Neurais de Kohonen Unidimensional foram aplicados a cada base de dados por 10 vezes. Na comparação destes métodos sugere-se a utilização de mais repetições. Neste caso, sugere-se ainda a utilização de testes estatísticos para determinação do melhor método.

Para a definição do melhor método de agrupamento para os dados de instrumentação geotécnica-estrutural da Itaipu, sugere-se a aplicação da Simulação de Monte Carlo. Depois de ajustar um modelo às séries temporais, a partir dele, pode-se gerar séries de dados sintéticas variando a variância do ruído, e então, aplicam-se os métodos em todos os conjuntos de dados obtidos.

REFERÊNCIAS

ABNT – ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **Ações e Segurança nas Estruturas (NBR 8681)**, 2003.

AZZAG, H.; VENTURINI, G.; OLIVER, A.; GUINOT, C. A hierarchical ant based clustering algorithm and its use in three real-world applications. **European Journal of Operational Research**, v. 179, p. 906-922, 2007.

BOX, G.E.P.; JENKINS, G.M. **Time Series Analysis, forecasting and control.** Ed. Holden Day, 1976.

BRASIL. **Projeto de Lei Nº 1.181/2003**. Estabelece diretrizes para verificação da segurança de barragens de cursos de água para quaisquer fins e para aterros de contenção de resíduos líquidos industriais. Disponível em http://www.emtermos.com.br/ABMS/PL_1181.pdf. Acesso em 19/06/2009.

BORYCZKA, U. Finding groups in data: Cluster analysis with ants. **Applied Soft Computing**, v. 9, p. 61-70, 2009.

BOWLES, D.S.; ANDERSON, L.R.; GLOVER, T.F.; CHUHAN, S.S. Dam Safety Decision-Making: Combining Engineering Assessments With Risk Information, **Proc. of 2003 US Society on Dams Annual Lecture***,* 2003.

BUZZI, M. F. **Avaliação das Correlações de Séries Temporais de Leituras de Instrumentos de Monitoração Geotécnicos-Estrutural e Variáveis Ambientais em Barragens Estudo de Caso de Itaipu**. 101 f. Dissertação (Mestrado em Métodos Numéricos em Engenharia) – Setor de Ciências Exatas, Universidade Federal do Paraná, Curitiba, 2007.

CBGB - Comitê Brasileiro de Grandes Barragens. **Diretrizes para a inspeção e avaliação de segurança de barragens em operação.** Rio de Janeiro, 1983. 26 p.

DIBIAGIO, E. Question 78 - Monitoring of Dams and Their Foundations – General Report. **Proc. Of Twentieth Congress on Large Dams, ICOLD**, Beijing, p. 1459-1545, 2000.

DYMINSKI, A. S., STEINER, M. T. A.; VILLWOCK, R. Hierarchical Ordering of Extensometers Readings from Itaipu Dam. In: **First International Symposium on Life-Cycle Civil Engineering -IALCCE' 08**, Varenna – Itália, 2008.

DINIZ, C. A. R.; LOUZADA NETO, F. **Data mining: uma introdução.** São Paulo: ABE, 2000.

DINIZ, G. B.; BERLATO, M. A.; CHARKE, R. T.; FONTANNA, D. C. Identificação de regiões homogêneas de temperaturas máxima e mínima do Rio Grande do Sul. **Revista Brasileira de Agrometeorologia**, Santa Maria, v. 11, p. 303-312, 2003.

DORIGO, M.; BLUM, C. Ant colony optimization theory: A survey. **Theoritical Computer Science**, v. 344, p. 243-278, 2005.

DORIGO, M.; CARO, G. D.; GAMBARDELLA L. M. Ant algorithms for discrete optimization. **Artificial Life**, v. 5, p. 137-172, Belgium, 1999.

DORIGO, M.; GAMBARDELLA, L. M. Ant colonies for the traveling salesman problem. **BioSystems**, v. 43, n. 2, p. 73–81, 1997.

DORIGO, M.; STÜTZLE, T. **Ant colony optimization**. Cambridge: MIT Press, 2004.

DORIGO, M.; MANIEZZO, V.; COLORNI, A. Ant System: Optimization by a colony of cooperating agents. **IEEE Transactions on Systems, Man, and Cybernetics – Part B**, v. 26, n. 1, p. 1–26, 1996.

DORIGO, M.; BONABEAU, E.; THERAULAZ, G. Ant algorithms and stigmery. **Future Generation Computer Systems**, v. 16, n. 8, p. 851–871, 2000.

DUARTE, J. M. G.; CALCINA, A. M.; GALVÁN, V. R. Instrumentação geotécnica de Obras Hidrelétricas Brasileiras: Alguns Casos Práticos Atuais. In: XIII COBRAMSEG - Congresso Brasileiro de Mecânica dos Solos e Eng. Geotécnica, Curitiba, 2006. **XIII COBRAMSEG - Congresso Brasileiro de Mecânica dos Solos e Eng. Geotécnica**. Curitiba: ABMS, 2006. CD-ROM.

FAUSETT, L. **Fundamentals of Neural Networks – Architectures, Algorithms, and Applications**. New Jersey: Prentice Hall, 1994.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P.; UTHRUSAMY, R. **Advances in knowledge Discovery & Data Mining.** California: AAAI/MIT, 1996.

FEMA – Federal Emergency Management Agency, **Federal Guidelines For Dam Safety,** U. S. Department Of Homeland Security, USA , 2004.

FERNANDES, C.; MORA, A. M.; MERELO, J. J.; RAMOS, V.; LAREDO, J. L. J. kohonAnts: A Self-Organizing Ant Algorithm for Clustering and Pattern Classification. **Artificial Life**, v. 9, p. 428-435, 2008.

FREITAS, A. A. **Data Mining and Knowledge Discovery with Evolutionary Algorithms.** New York: Springer, 2002.

GHOSH, A.; HALDER, A.; KOTHARI, M.; GHOSH, S. Aggregation pheromone density based data clustering. **Information Sciences**, v. 178, p. 2816-2831, 2008.

GAVRILOV, M; ANGUELOV, D.; INDYK, P.; MOTWANI, R. Mining the stock market: Which measure is best? In: ACM Int'l Conference on Knowledge Discovery And Data Mining, 2000. **Proc. of KDD'00**: 2000. p. 487-496.

HAIR JR, J.F.; ANDERSON, R.E.; TATHAM, R.L.; BLACK, W.C. **Análise Multivariada de Dados**. Tradução de: SANTANNA, A. S.; CHAVES NETO, A. Porto Alegre: Bookman, 2005.

HANDL, J.; KNOWLES, J.; DORIGO, M. Ant-Based Clustering and Topographic Mapping. **Artificial Life**, v. 12, n. 1, p. 35-61, 2006.

HANDL,J.; MEYER, B. Ant-based and swarm-based clustering. **Swarm Intell**, v. 1, p. 95-113, 2007.

HARRALD, J. R.; RENDA-TANALI, I.; SHAW, G.L.; RUBIN, C.B.; YELETAYSI, S. **Review of Risk Based Prioritization/Decision Making Methodologies for Dams**. Technical Report, US Army Corps of Engineers, 2004.

HAYKIN, S. **Redes neurais: princípios e prática.** Tradução: Paulo Martins Engel. Porto Alegre: Bookman, 2001.

HO, Y. C.; PEPUNE, D. L. Simple Explanation of the No-Free-Lunch Theorem and Its Implications. **Journal of Optimization Theory and Applications**, v. 115, n. 3, p. 549-570, 2002.

ICOLD - **International Commission on Large Dams**. http://www.icold-cigb.org, 2008.

ITAIPU – Itaipu Binacional. Disponível em http://www.itaipu.gov.br/. Acesso em 28/08/2008.

JAIN A. K., MURTY M. N., FLYNN P. J. Data clustering: a review. **ACM Computing Surveys**. v. 31, n. 3, 1999.

JOHNSON, R.A.; WICHERN, D.W. **Applied Multivariate Statistical Analysis.** Fourth Edition. New Jersey: Prentice Hall, 1998.

KALUSTYAN, E. S. Assessment and role of risk in dam building. **Hydrotechnical Construction**, v. 33, n. 12, 1999.

KEOGH, E.; XI, X.; WEI, L. & RATANAMAHATANA, C. A. **The UCR Time Series Classification/Clustering** Homepage: www.cs.ucr.edu/~eamonn/time_series_data/, 2006.

KOHONEN, T. **Self-Organizing Map**. Berlin: Springer-Verlag, 1995.

KUO, R. J.; WANG, H. S.; HU, T-L; CHOU, S.H. Application of Ant K-Means on Clustering Analysis. **Computers and Mathematics with Applications**, v. 50, p. 1709-1724, 2005.

KRÜGER, C. M. **Análise de Confiabilidade Estrutural Aplicada às Barragens de Concreto**. 157 f. Tese (Doutorado em Métodos Numéricos em Engenharia) – Setor de Ciências Exatas, Universidade Federal do Paraná, Curitiba, 2008.

LIAO, T. W. Clustering of time series data – a survey. **Pattern Recognition**. v. 38, p. 1857-1874, 2005.

MATLAB R2008b – The MatWorks, *MATLAB* (R2008b), The MathWorks Inc., Natick, 2008.

MATOS, S. F. **Avaliação de Instrumentos para Auscultação de Barragem de Concreto. Estudo de caso: Deformímetros e Tensômetros para Concreto na Barragem de Itaipu**. 106 f. Dissertação (Mestrado em Construção Civil). – Setor de Ciências Exatas, Universidade Federal do Paraná, Curitiba, 2002.

MENESCAL, R. de A. **Gestão da Segurança de Barragens no Brasil - Proposta de um Sistema Integrado, Descentralizado, Transparente e Participativo**. 769 f. Tese (Doutorado em Engenharia Civil) - Departamento de Engenharia Hidráulica e Ambiental, Universidade Federal do Ceará, Fortaleza, 2009.

MORETTIN, P. A.; TOLOI, C. M. DE C. **Previsão de Séries Temporais**, São Paulo: Atual, 1985.

OSAKO, C. I. **A Manutenção dos Drenos nas Fundações de Barragens - O Caso da Usina Hidrelétrica de Itaipu.** 126 f. Dissertação (Mestrado em Construção Civil). – Setor de Ciências Exatas, Universidade Federal do Paraná, Curitiba, 2002.

OCA, M. A. M. de; GARRIDO, L.; AGUIRRE, J. L. Efectos de la comunicación directa entre agentes em los algoritmos de agrupación de clases basados em el comportamiento de insectos sociales. **Inteligência Artificial, Revista Iberoamericana de Inteligência Artificial**, n. 25, p. 59-69, 2005.

SANCHEZ, P. F. **Mapeamento Espaço-Temporal e Previsão de Pressões Piezométricas em Maciços Rochosos de Fundações de Grandes Barragens – Estudo de Caso de Itaipu**. Dissertação (Mestrado em Construção Civil). – Setor de Ciências Exatas, Universidade Federal do Paraná, Curitiba, 2009.

SARÉ, A. R.; LIGOCKI, L. P.; SAYÃO, A.; GERSCOVICH, D. M.; PINHEIRO, G. F. Revisão das Condições de Segurança da Barragem de Curuá-Una (PA). In: XIII COBRAMSEG - Congresso Brasileiro de Mecânica dos Solos e Eng. Geotécnica, Curitiba, 2006. **XIII COBRAMSEG - Congresso Brasileiro de Mecânica dos Solos e Eng. Geotécnica**. Curitiba: ABMS, 2006. CD-ROM.

SILVA NETO, M. A.; VILLWOCK, R.; STEINER, M. T. A.; DYMINSKI, A. S.; SCHEER, S. Mineração Visual de Dados Aplicada à Extração do Conhecimento nos Dados de Instrumentação da Barragem de Itaipu. In: XL Simpósio Brasileiro de Pesquisa Operacional - SBPO, João Pessoa, 2008. **XL Simpósio Brasileiro de Pesquisa Operacional – SBPO**. João Pessoa: SOBRAPO, 2008.

SILVEIRA, J. F. A. **Instrumenatação e Comportamento de Fundações de Barragens de Concreto.** São Paulo: Oficina de Textos, 2003.

SILVER, D.L. Knowledge Discovery and Data Mining. **Technical Report MBA6522 CogNova Technologies London Health Science Center**, 1996.

SIQUEIRA, P. H. **Uma Nova Abordagem na Resolução do Problema do Caixeiro Viajante**. 116 f. Tese (Doutorado em Métodos Numéricos em Engenharia) – Setor de Ciências Exatas, Universidade Federal do Paraná, Curitiba, 2005.

SOCHA, k.; DORIGO, M. Ant colony optimization for continuous domains. **EuropeanJournal of Operational Research**, v. 185, p. 1155-1173, 2008.

STATGRAPHICS PLUS 5.1 – *Statgraphics Plus* 5.1, Statistical Graphics Corp., Rockville, 2001.

STEINBACH, M.; TAN, P. N.; KUMAR, V.; POTTER, C.; KLOOSTER, S. Data Mining for the Discovery of Ocean Climate Indices. In: Workshop on Mining Scientific Datasets, Second SIAM International Conference on Data Mining, 2002. In: Fifth Workshop on Scientific Data Mining, 2002. **Proc of the Fifth Workshop on Scientific Data Mining**. 2002.

TAN, P. N.; STEINBACH, M.; KUMAR, V. **Introduction to Data Mining**. Inc. Boston, MA, USA: Addison-Wesley Longman Publishing Co., 2005.

TAN, S. C.; TING, K. M.; TENG, S. W. Examining Dissimilarity Scaling in Ant Colony Approaches to Data Clustering. In: ACAL, 2007. **ACAL 2007**. Springer-Verlag, 2007.

U.S. Army Corps of Engineers. **Instrumentation for Concrete Structures. Engineering and Design**. Engineer Manual Nº No. 1110-2-4300, Washington, DC, 1987.

U.S. Army Corps of Engineers. **Instrumentation of embankment dams and levees. Engineering and Design**. Engineer Manual Nº 1110-2-1908, Washington, DC, 1995.

VILLWOCK, R.; STEINER, M. T. A. Agrupamento baseado em Colônia de Formigas: Estudo Comparativo de Algoritmos para Recuperação dos Grupos. In: XII Encontro Regional de Matemática Aplicada e Computacional, Foz do Iguaçú, 2008. **XII Encontro Regional de Matemática Aplicada e Computacional**. Foz do Iguaçú: 2008. CD-ROM.

VILLWOCK, R.; STEINER, M. T. A. Análise do Desempenho do Algoritmo de Agrupamento Baseado em Colônia de Formigas Modificado. In: XXXII Congresso Nacional de Matemática Aplicada e Computacional, Cuiabá, 2009. **XXXII Congresso Nacional de Matemática Aplicada e Computacional**. Cuiabá: SBMAC, 2009a. CD-ROM.

VILLWOCK, R.; STEINER, M. T. A. Análise do Desempenho de um Algoritmo de Agrupamento Modificado Baseado em Colônia de Formigas. In: XLI Simpósio Brasileiro de Pesquisa Operacional, Porto Seguro, 2009. **XLI Simpósio Brasileiro de Pesquisa Operacional**. Porto Seguro: SOBRAPO, 2009b. CD-ROM.

VILLWOCK, R.; STEINER, M. T. A.; DYMINSKI, A. S. Data Mining Applied to the Instrumentation Data Analysis of a Large Dam. In: Seventh International Conference on Intelligent Systems Design and Applications - ISDA, Rio de Janeiro, 2007. **Seventh International Conference on Intelligent Systems Design and Applications – ISDA**. Rio de Janeiro: IEEE, 2007.

VILLWOCK, R.; STEINER, M. T. A.; DYMINSKI, A. S.; CHAVES NETO, A. Análise Multivariada Aplicada à Análise de Dados de Instrumentação de Barragens – Estudo

de Caso de Itaipu. In: XIV Simpósio de Engenharia de Produção - SIMPEP, Bauru, 2007. **Anais do XIV Simpósio de Engenharia de Produção – SIMPEP**. Bauru: 2007.

VILLWOCK, R.; STEINER, M. T. A.; DYMINSKI, A. S.; CHAVES NETO, A. Data Mining Applied to the Instrumentation Analysis of a Large Dam. In: PONCE, J.; KAROHOCA, A. **Data Mining and Knowledge Discovery in Real Life Applications**. Vienna: In-Tech, 2009. p. 389-406.

VIZINE, A. L.; DE CASTRO, L. N.; HRUSCHKA, E. R.; GUDWIN, R. R. Towards improving clustering ants: an adaptive ant clustering algorithm. **Informatica**, v. *29*, p. 143–154, 2005.

WITTEN, I. H.; FRANK, E. **Data Mining: Pratical Machine Learning Tools and Techniques with Java Implementations.** San Francisco: Morgan Kaufmann Publishers, 2000.

YANG, Y.; KAMEL, M. S. An aggregated clustering approach using multi-ant colonies algorithms. **Pattern Recognition**, v. 39, 1278-1289, 2006.

YENIGUN, K.; ERKEK, C. Reliability in dams and the effects of spillway dimensions on risk levels. **Water Resources Management**, v. 21, p. 747-760, 2007.

ANEXOS

ANEXO 1
INFORMAÇÕES SOBRE AS BASES DE DADOS UTILIZADAS

A base de dados IRIS é composta de 150 padrões (flores). Nesta base de dados são conhecidos os grupos a que cada flor pertence. Os 150 padrões são divididos em três grupos com 50 padrões em cada grupo: Íris Setosa, Íris Versicolour e Íris Virginica. Cada padrão é constituído por quatro atributos numéricos: comprimento da pétala, largura da pétala, comprimento da sépala e largura da sépala.

A base de dados WINE é composta de 178 padrões (vinhos). Nesta base de dados também são conhecidos os grupos a que cada padrão pertence. Os 178 padrões são divididos em três grupos: 59 padrões pertencem ao grupo 1, 71 padrões pertencem ao grupo 2 e 48 padrões pertencem ao grupo 3. Cada padrão é constituído de 13 atributos numéricos. Os atributos são resultados de uma análise química.

A base de dados PIMA Indians Diabetes é composta de 768 padrões (mulheres menores de 21 anos descendentes da tribo Pima). Nesta base de dados também são conhecidos os grupos a que cada padrão pertence. Os 768 padrões são divididos em dois grupos: 500 padrões pertencem ao grupo 1 (não tem diabetes) e 268 padrões pertencem ao grupo 2 (tem diabetes). Cada padrão é constituído de 8 atributos numéricos. Os atributos são índice de massa corpórea, idade, alguns resultados de exames laboratoriais, entre outros.

A base de dados GUN (arma) é composta de 200 padrões (séries temporais). Foram utilizados vídeos de vigilância com atores femininos e masculinos. Em cada vídeo foram capturadas 30 seções (puxa a arma e aponta) e a posição da mão direita dos atores foi obtida a cada 5 segundos. Nesta base de dados também são conhecidos os grupos a que cada padrão pertence. Os 200 padrões são divididos em dois grupos: 100 padrões pertencem ao grupo 1 (puxar a arma) e 100 padrões pertencem ao grupo 2 (apontar a arma). Cada padrão é constituído de 150 atributos numéricos. Os atributos são as posições da mão direita dos atores.

Não foram encontradas informações sobre a base de dados LIGHTNING-2.

ANEXO 2
A USINA HIDRELÉTRICA DE ITAIPU

A Itaipu Binacional, maior hidrelétrica em produção de energia do mundo, teve o início da sua construção em 1973 em um trecho do Rio Paraná conhecido por Itaipu que, em tupi, quer dizer “a pedra que canta”, localizado no coração da América do Sul na divisa entre o Paraguai e o Brasil (ITAIPU, 2008). Em 1982, chegaram ao fim as obras da barragem, sendo a última unidade geradora inaugurada em 2008.

Atualmente, a barragem de Itaipu possui 20 unidades geradoras de 700 MW (*megawatts*) cada, gerando uma potência total instalada de 14.000 MW. No ano 2000, a Itaipu Binacional bateu seu recorde em geração de energia, cerca de 93,4 bilhões de quilowatts-hora (KWh). É responsável pelo abastecimento de 95% da energia elétrica consumida no Paraguai e 24% de toda a demanda do mercado brasileiro.

A figura 1, a seguir, mostra a estrutura geral da barragem de Itaipu e o quadro 1 apresenta as principais características dos trechos da barragem apontados na referida figura 1.

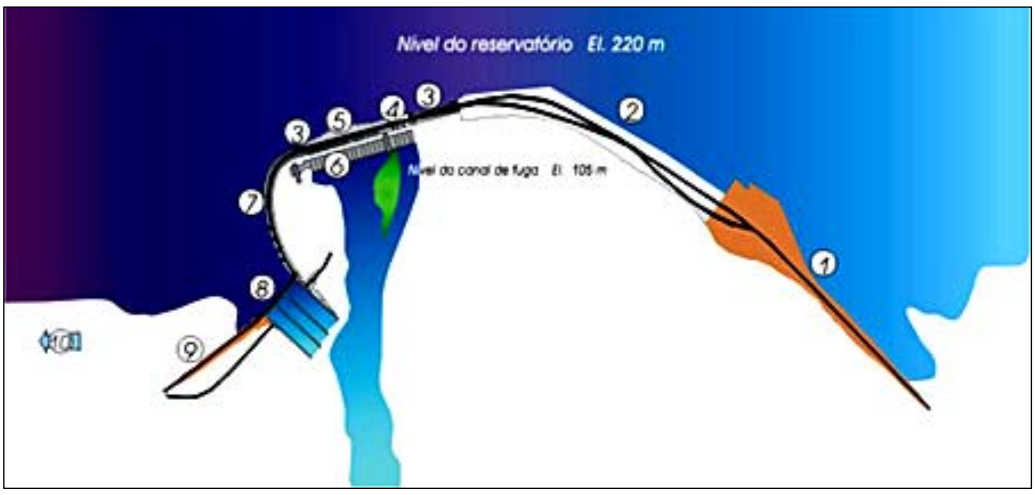

Figura 1 – Estrutura geral do complexo Itaipu (ITAIPU, 2008).

Nos quadros 1 e 2, podem-se observar os tipos e quantidades de instrumentos instalados, respectivamente, na fundação e no concreto dos blocos do trecho F da barragem.

Trecho		Estrutura	Comprimento (m)	Altura Máxima (m)
1	**Barragem Auxiliar**	Terra	2294	30
2	**Barragem Auxiliar**	Enrocamento	1984	70
3 e 7	**Barragens Laterais**	Contraforte	1438	81
4	**Estrutura de Desvio**	Concreto Maciço	170	162
5	**Barragem Principal (Trecho F)**	Gravidade Aliviada	612	196
9	**Barragem Auxiliar**	Terra	872	25
Outros trechos		**Características**		
6	**Casa de Força**	20 Unidades Geradoras		
8	**Vertedouro**	350 m de Largura		

Quadro 1 – Características dos trechos da Itaipu.

Instrumento	Sigla	Blocos do Trecho F					Total
		5/6	13/14	15/16	19/20	35/36	
Rosetas Deformímetro	**RD**	4	-	-	11	-	15
Tensômetro	**TN**	1	-	-	4	-	5
Rosetas de Tensômetro	**RT**	2	-	-	6	-	8
Medidor de Junta Interna	**JM**	-	-	-	7	-	7
Pêndulo Direto	**PD**	5	6	-	6	4	21
Pêndulo Invertido	**PI**	3	1	1	1	-	6
Termômetro na Massa	**TM**	3	-	-	17	3	23
Termômetro na Superfície	**TS**	2	-	-	6	2	10
Total por Bloco		20	7	1	58	9	95

Quadro 2 – Quantidades e tipos de instrumentos no concreto encontrados nos blocos do trecho F da Itaipu (ITAIPU, 2008).

A figura 2 mostra os derrames de “A” a “E” do perfil basáltico do maciço de fundação da Itaipu.

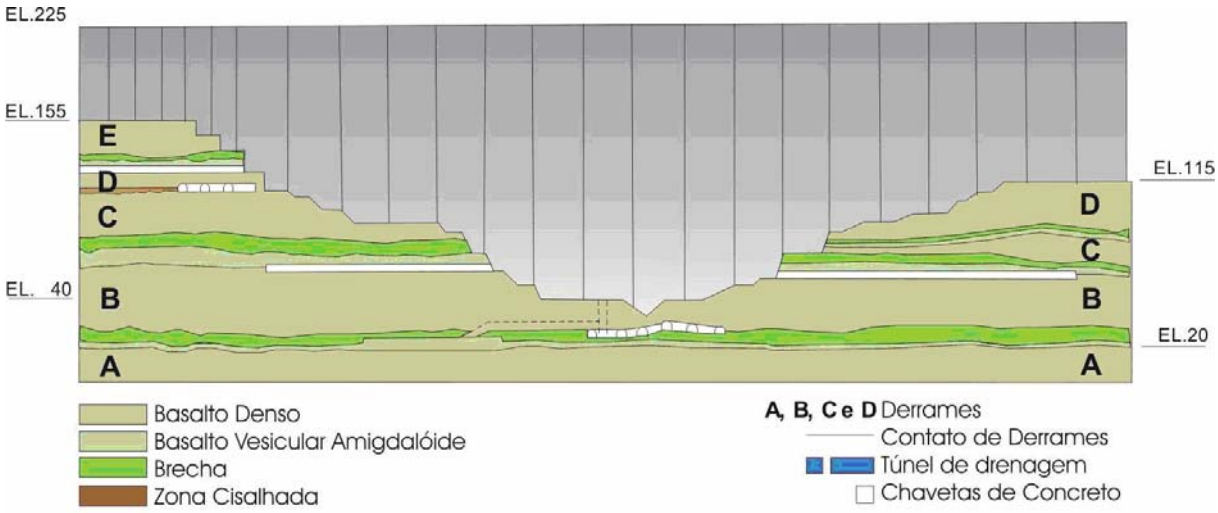

Figura 2 – Perfil basáltico do maciço de fundação da Itaipu (ITAIPU, 2008).

Instrumento	Sigla	Blocos do Trecho F																Total
		1/2	3/4	5/6	7/8	9/10	11/12	13/14	15/16	17/18	19/20	21/22	23/24	27/28	29/30	31/32	35/36	
Piezômetro Standpipe	OS	-	4	6	5	-	6	7	3	6	8	-	4	10	-	6	9	74
Piezômetro Geomor	PG	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0
Extensômetro de Haste	EM	4	-	1	-	-	-	3	5	4	4	1	-	4	-	-	4	30
Medidor de Aterro	MA	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0
Medidor Triortogonal	MT	-	1	-	-	1	4	1	1	-	-	1	1	-	-	1	-	11
Célula de Pressão Total	CL	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0
Medidor de Vazão	MV	-	1	-	-	-	2	-	2	-	-	2	1	-	1	-	-	9
Drenos	DR	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0
Medidor de Nível d'Água	PZ	-	-	-	-	-	-	1	1	1	1	-	-	-	-	-	-	4
Total por Bloco		4	6	7	5	1	12	12	12	11	13	4	6	14	1	7	13	128

Quadro 3 – Quantidades e tipos de instrumentos na fundação encontrados nos blocos do trecho F da Itaipu (ITAIPU, 2008).

ANEXO 3

EXEMPLO ACADÊMICO DO FUNCIONAMENTO DO ALGORITMO DE AGRUPAMENTO BASEADO EM FORMIGAS

A matriz de dados *A,* com dimensão 5x6, tem valores dentro do intervalo [0,1]. São cinco padrões (linhas) e cada padrão tem seis atributos (colunas). É apresentada, a seguir, uma iteração do algoritmo de Agrupamento Baseado em Formigas.

$$A = \begin{bmatrix} 0,43 & 0,18 & 0,58 & 0,09 & 0,69 & 0,39 \\ 0,66 & 0,03 & 0,18 & 0,83 & 0,85 & 0,69 \\ 0,13 & 0,27 & 0,11 & 0,33 & 0,29 & 0,71 \\ 0,12 & 0,32 & 0,13 & 0,29 & 0,25 & 0,81 \\ 0,19 & 0,72 & 0,05 & 0,62 & 0,57 & 0,67 \end{bmatrix}$$

A matriz *D* é a matriz das distâncias Euclidianas entre os padrões. A matriz *DN* é a matriz das distâncias Euclidianas padronizadas no intervalo [0,1].

$$D = \begin{bmatrix} 0 & 0,95 & 0,8 & 0,85 & 1 \\ 0,95 & 0 & 0,95 & 1,02 & 0,91 \\ 0,8 & 0,95 & 0 & 0,13 & 0,61 \\ 0,85 & 1,02 & 0,13 & 0 & 0,63 \\ 1 & 0,91 & 0,61 & 0,63 & 0 \end{bmatrix} \quad DN = \begin{bmatrix} 0 & 0,93 & 0,78 & 0,84 & 0,98 \\ 0,93 & 0 & 0,93 & 1 & 0,9 \\ 0,78 & 0,93 & 0 & 0,12 & 0,6 \\ 0,84 & 1 & 0,12 & 0 & 0,62 \\ 0,98 & 0,9 & 0,6 & 0,62 & 0 \end{bmatrix}$$

Será utilizada uma grade bidimensional 4x4, totalizando 16 células. Neste exemplo foram convencionados: número de formigas igual a “3”, raio de vizinhança igual a “1”, comprimento do passo *L* igual a “2”, *α* igual a 0,5, o valor mínimo para a probabilidade de carregar um padrão igual a 0,5 e o valor mínimo para a probabilidade de descarregar um padrão igual a 0,5.

Foi realizada a distribuição dos padrões sobre a grade. A cada formiga foi associado um padrão. A distribuição das formigas e dos padrões na grade pode ser observada na figura 1.

Por se tratar de exemplo didático, as formigas serão selecionadas sequencialmente. A formiga *F1* está associada ao padrão *P1* e se encontra na posição 2. Executando um passo de comprimento 2 para direita a posição atual é a

posição 4. Na avaliação da função *f*, considerando o raio de vizinhança igual a “1”, o único padrão vizinho ao padrão *P1* é o padrão *P3*. O valor da função *f* é “0”. A probabilidade da formiga *F1* descarregar o padrão *P1* na posição 4 é “0”, ou seja, o padrão *P1* não deve ser descarregado.

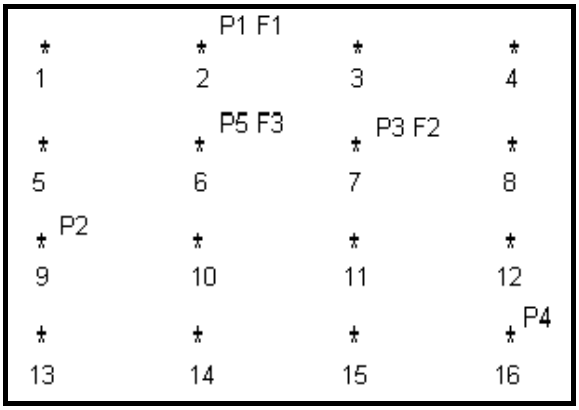

Figura 1 – Distribuição das formigas e dos padrões na grade 1 – EXEMPLO.

A formiga *F2* está associada ao padrão *P3* e se encontra na posição 7. Executando um passo de comprimento dois para baixo a posição atual é a posição 15. Na avaliação da função *f*, considerando o raio de vizinhança igual a “1”, o único padrão vizinho ao padrão *P3* é o padrão *P4*. O valor da função *f* é “0,76”. A probabilidade da formiga *F2* descarregar o padrão *P3* na posição 15 é “0,51”, ou seja, o padrão *P3* deve ser descarregado na posição 15.

Como a formiga *F2* descarregou seu padrão, outro padrão deve ser carregado. Os padrões livres são os padrões *P2* e *P4*. Iniciando com o padrão *P2*, na avaliação da função *f*, considerando o raio de vizinhança igual a “1”, o único padrão vizinho ao padrão *P2* é o padrão *P5*. O valor da função *f* é “0”. A probabilidade da formiga *F2* carregar o padrão *P2* é “1”, ou seja, o padrão *P2* deve ser carregado pela formiga *F2*. A nova distribuição na grade pode ser observada na figura 2.

A formiga *F3* está associada ao padrão *P5* e se encontra na posição 6. Executando um passo de comprimento dois para esquerda a posição atual é a posição 5 (como não é possível realizar o passo de comprimento dois, a posição foi definida considerando o maior passo possível nesta direção). Na avaliação da função *f*, considerando o raio de vizinhança igual a “1”, os padrões vizinhos ao padrão *P5* são os padrões *P1* e *P2*. O valor da função *f* é “0”. A probabilidade da

formiga *F3* descarregar o padrão *P5* na posição 5 é “0”, ou seja, o padrão *P5* não deve ser descarregado.

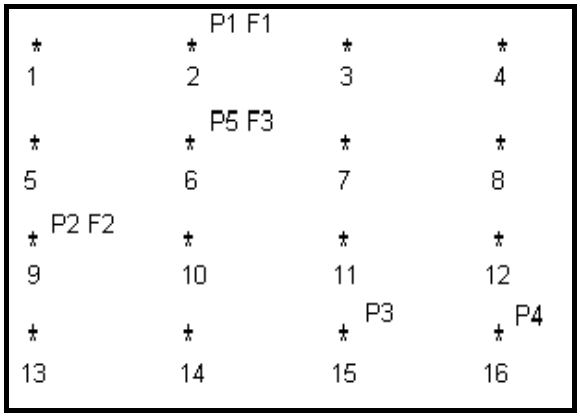

Figura 2 – Distribuição das formigas e dos padrões na grade 2 – EXEMPLO.

Como todas as formigas foram selecionadas, termina a primeira iteração. A solução para a primeira iteração pode ser observada na figura 3.

Supondo que esta seja a solução final, aplicamos o algoritmo de recuperação do agrupamento. Inicia-se admitindo que cada padrão forma um grupo, portanto o grupo *G1* é formado pelo padrão *P1*, o grupo *G2* é formado pelo padrão *P2*, o grupo *G3* é formado pelo padrão *P3*, o grupo *G4* é formado pelo padrão *P4* e o grupo *G5* é formado pelo padrão *P5*. Depois, calcula-se a distância Euclidiana entre os grupos, baseando-se nas distâncias das posições de seus elementos na grade. A matriz destas distâncias é a matriz *DG*.

$$DG = \begin{bmatrix} 0 & 2,24 & 3,16 & 3,61 & 1 \\ 2,24 & 0 & 2,24 & 3,16 & 1,41 \\ 3,16 & 2,24 & 0 & 1 & 2,24 \\ 3,61 & 3,16 & 1 & 0 & 2,83 \\ 1 & 1,41 & 2,24 & 2,83 & 0 \end{bmatrix}$$

A menor distância é “1”, observada entre os grupos *G1* e *G5* e entre os grupos *G3* e *G4*. Escolheu-se agrupar primeiro os grupos *G1* e *G5* e este será denominado grupo *G1*. Portanto, os grupos agora são: grupo *G1*, formado pelos padrões *P1* e *P5*; grupo *G2*, formado pelo padrão *P2*; grupo *G3*, formado pelo padrão *P3* e o grupo *G4*, formado pelo padrão *P4*. Recalculando a matriz *DG*, tem-se:

$$DG = \begin{bmatrix} 0 & 1,41 & 2,24 & 2,83 \\ 1,41 & 0 & 2,24 & 3,16 \\ 2,24 & 2,24 & 0 & 1 \\ 2,83 & 3,16 & 1 & 0 \end{bmatrix}$$

A menor distância é “1”, observada entre os grupos *G3* e *G4*. Agrupou-se os grupos *G3* e *G4* e este será denominado grupo *G3*. Os grupos agora são: grupo *G1*, formado pelos padrões *P1* e *P5*; grupo *G2*, formado pelo padrão *P2*; grupo *G3*, formado pelos padrões *P3* e *P4*. Recalculando a matriz DG, tem-se:

$$DG = \begin{bmatrix} 0 & 1,41 & 2,24 \\ 1,41 & 0 & 2,24 \\ 2,24 & 2,24 & 0 \end{bmatrix}$$

A menor distância é “1,41”, observada entre os grupos *G1* e *G2*. Agrupou-se os grupos *G1* e *G2* e este será denominado grupo *G1*. Os grupos agora são: grupo *G1*, formado pelos padrões *P1*, *P2* e *P5* e grupo *G2*, formado pelos padrões *P3* e *P4*. Recalculando a matriz DG, tem-se:

$$DG = \begin{bmatrix} 0 & 2,24 \\ 2,24 & 0 \end{bmatrix}$$

Agrupou-se os grupos *G1* e *G2* e este será denominado grupo *G1*. O grupo *G1* agora é formado pelos padrões *P1*, *P2*, *P3*, *P4* e *P5*. A figura 3 mostra o dendrograma.

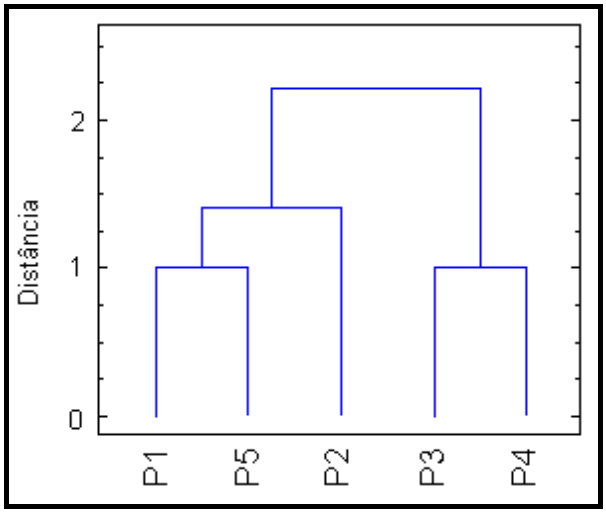

Figura 3 – Dendrograma – EXEMPLO.